


# Learning Self-distilled Features for Facial Deepfake Detection Using Visual Foundation Models: General Results and Demographic Analysis


Yan Martins Braz Gurevitz Cunha   [ Telemidia Lab. – Pontifical Catholic University of Rio de Janeiro | [yangurevitz@telemidia.puc-rio.br](mailto:yangurevitz@telemidia.puc-rio.br) ]


Bruno Rocha Gomes  [ Telemidia Lab. – Pontifical Catholic University of Rio de Janeiro | [brunogomes@telemidia.puc-rio.br](mailto:brunogomes@telemidia.puc-rio.br) ]


José Matheus C. Boaro  [ Telemidia Lab. – Pontifical Catholic University of Rio de Janeiro | [boaro@telemidia.puc-rio.br](mailto:boaro@telemidia.puc-rio.br) ]

Daniel de Sousa Moraes  [ Telemidia Lab. – Pontifical Catholic University of Rio de Janeiro | [danielmoraes@telemidia.puc-rio.br](mailto:danielmoraes@telemidia.puc-rio.br) ]

Antonio José Grandson Busson  [ BTG Pactual | [antonio.busson@btgpactual.com](mailto:antonio.busson@btgpactual.com) ]

Julio Cesar Duarte  [ Military Institute of Engineering | [duarte@ime.br](mailto:duarte@ime.br) ]

Sérgio Colcher  [ Telemidia Lab. – Pontifical Catholic University of Rio de Janeiro | [colcher@inf.puc-rio.br](mailto:colcher@inf.puc-rio.br) ]

 *Telemidia Lab. – Departamento de Informática – PUC-Rio, Av. Marquês de São Vicente 225, Gávea, Rio de Janeiro – RJ, 22453-900, Brazil.*

**Received:** 10 February 2024 • **Accepted:** 26 June 2024 • **Published:** 09 July 2024

**Abstract:** Modern deepfake techniques produce highly realistic false media content with the potential for spreading harmful information, including fake news and incitements to violence. Deepfake detection methods aim to identify and counteract such content by employing machine learning algorithms, focusing mainly on detecting the presence of manipulation using spatial and temporal features. These methods often utilize Foundation Models trained on extensive unlabeled data through self-supervised approaches. This work extends previous research on deepfake detection, focusing on the effectiveness of these models while also considering biases, particularly concerning age, gender, and ethnicity, for ethical analysis. Experiments with DINOv2, a novel Vision Transformer-based Foundation Model, trained using the diverse Deepfake Detection Challenge Dataset, which encompasses several lighting conditions, resolutions, and demographic attributes, demonstrated improved deepfake detection when combined with a CNN classifier, with minimal bias towards these demographic characteristics.

**Keywords:** Deepfake Detection, Foundation Models, Machine Learning, Demographic Analysis, Self-Supervised Methods

## 1 Introduction

There has been a surge in both the creation and consumption of content from social media platforms. Sites and applications such as Facebook, Instagram, and TikTok actively encourage the culture of public sharing, urging individuals to post photos and videos, not only about their personal lives but also about their relatives and friends, rewarding these actions with increased visibility and engagement [Almond Solutions, 2021]. However, this seemingly naive action may have undesirable consequences when considering the potential ramifications of such widespread sharing since this content becomes susceptible to exploitation through deepfake techniques [Kiefer, 2023]. Consequences range from misinformation and loss of trust to even identity theft, endangering essential sectors of society such as journalism, politics, and entertainment.

Deepfake techniques create almost realistic synthetic deceptive media through artificial intelligence algorithms, raising serious concerns across several domains. One recent case of deepfake misuse was the spread of explicit fake generated images of the singer Taylor Swift, which forced platforms like X (formerly Twitter) to take measures to turn off related searches of her content temporarily [Schmunk, 2024] and renewed discussions in countries such as the United States to legally ban the creation and distribution of this kind of content [Beaumont-Thomas, 2024].

Modern deepfake methods are known for their capacity to produce realistically adulterated or falsified multimedia content that can potentially spread harmful and malicious information, disseminate fake news, or even promote violence and hate. Initially, this technology became popular through small machine learning communities, which resulted in easy-to-use, open-source implementations such as FakeApp, DFaker, Faceswap, and DeepFaceLab [Li *et al.*, 2022]. In these tools, the face of a target person is replaced by one of another person, maintaining the expressions and poses of the target face, which can facilitate the dissemination of fake news and malicious content. Currently, modern deepfake generation methods can produce fake videos and images with realistic aspects by using Generative Adversarial Networks (GANs) [Creswell *et al.*, 2018], Variational Autoencoders [Khalid and Woo, 2020], and Diffusion Models [Corvi *et al.*, 2023].

Using such sophisticated techniques highlights the significant importance of the constant advances in refining deepfake detection techniques to effectively and immediately purge deceptive content from social media platforms.

Deepfake detection deals with designing methodologies that identify and mitigate the presence of deepfake content in digital media, like audio, photos, and videos, distinguishing authentic content from manipulated or synthesized media. This process usually involves using advanced machine learning algorithms that analyze subtle inconsistencies and

anomalies that may indicate the presence of manipulation, and its main approaches focus on the utilization of spatial or temporal features [Xu *et al.*, 2024].

To combat the spread of deepfake content and its harms, several large-scale datasets and models for deepfake detection have emerged in recent years, especially after Kaggle’s Deepfake Detection Challenge Dataset (DFDC) competition.<sup>1</sup> CNN-based models have generally shown good performance for this task, but the use of Vision Transformers (ViT) combined with CNN models were the ones that recently reached state-of-the-art [Heo *et al.*, 2021]. Despite this, the spatial locality of CNNs is still significant in discovering anomalies in images, being helpful in this task.

Recently, a new AI paradigm called *Foundation Model* (FM) has emerged. FMs comprise any deep learning model trained on massive amounts of unlabeled data, usually by self-supervised learning [Bommasani *et al.*, 2021]. Self-supervised features generated by pre-trained FMs can be used for various downstream tasks.

Although achieving optimal performance is a common objective in automated detection methods, studying how the algorithm behaves based on different input instances is also an important aspect of methodology design. From the inner details of a dataset to the nuances of algorithm flows, bias is a common issue studied in any artificial intelligence-derived system, as it may render them less reliable and more susceptible to false positives or negatives in certain specific circumstances. By understanding when bias becomes a problem in this context, it is possible to overcome the potential limitations of such detection methods and refine them to produce a more accurate classifier.

Gomes *et al.* [2023] focused on establishing a novel approach to detecting facial deepfakes by combining DINO (DIstillation with NO labels) [Caron *et al.*, 2021], a foundation model based on Vision Transformer that produces universal self-supervised features suitable for image-level visual tasks with traditional CNN-based classifiers. Those features were combined with the raw RGB channels to feed different CNN classifiers. By evaluating them over DFDC [Dolhan-sky *et al.*, 2020a], we showed improvements in deepfake facial detection in scenarios with different baselines. However, our original work did not differentiate between demographic groups or account for the impact of demographic attributes on performance, as shown in other deepfake detection methods [Xu *et al.*, 2024; Trinh and Liu, 2021]. Therefore, verifying how this approach’s performance changes across multiple demographic groups is important.

Especially concerning age, gender, and ethnicity, it is mandatory to study the implications of bias in artificial intelligence and machine learning for ethical considerations. Understanding how these algorithms may exhibit biases related to such sensitive attributes can contribute to developing fair and transparent models while ensuring that they do not perpetuate or increase societal inequities, reinforcing inclusivity.

With this perspective in mind, the present study extends upon the findings of Gomes *et al.* [2023] by making use of

an expanded dataset to further validate our original results, applying a new self-supervised method in DINOv2 [Oquab *et al.*, 2024] and investigating the following research question: *How do demographic classes of attributes behave within a self-supervised strategy aimed at classifying realistic facial deepfakes?*

The remainder of this paper is organized as follows. Section 2 describes some related work on facial deepfake generation and classification and studies on demographic bias and fairness in deepfake detection. Next, in Section 3, we introduce our proposed method for facial deepfake detection. Section 4 covers our expanded dataset and the experiments used to validate our original results, followed by Section 5, where we describe and analyze how our results are impacted by demographic features, investigating the fairness of our approach. Finally, Section 6 is devoted to our final remarks and conclusions.

## 2 Related Work

This section describes related work in the deepfake context. First, in section 2.1, we present works focused on deepfake generation. Next, we present recent works on deepfake detection in section 2.2. Finally, in section 2.3, we present works focused on bias and fairness analysis in deepfake detection.

### 2.1 Deepfake Generation

Perov *et al.* [2023] proposed DeepFaceLab, a deepfake framework for face-swapping tasks. This model is composed of three modules: (1) face detection using the S3FD Zhang *et al.* [2017] model; (2) extraction of facial landmarks using the 2DFAN [Bulat and Tzimiropoulos, 2017] and PR-Net [Feng *et al.*, 2018] algorithms; and finally, (3) segmentation of faces through the XSeg and TerausNet [Igloukov and Shvets, 2018] networks. DeepFaceLab was compared with the deepfakes<sup>2</sup> and Nirkin *et al.* [2018] models using the FaceForensics++ dataset [Rossler *et al.*, 2019] and the identical training setups. In experiments, models were evaluated using the following metrics: the accuracy of the head pose, the accuracy of facial expressions, the score obtained by segmentation masks via SSIM [Wang *et al.*, 2004], the perceptual loss [Johnson *et al.*, 2016]; and finally, the accuracy obtained by face verification using DLib [King, 2009]. The results obtained by DeepFaceLab were the best in all metrics except for the facial expressions metric.

Choi *et al.* [2018] proposed StarGAN, a model capable of performing image-to-image translation for multiple domains. The architecture was adapted from CycleGAN [Zhu *et al.*, 2017], containing a generator network composed of two convolution layers, six residual blocks, two transposed convolution and instance normalization layers, and a discriminator network based on PatchGANs [Isola *et al.*, 2017]. The training was performed on the CelebA [Liu *et al.*, 2015] and RaFD [Langner *et al.*, 2010] datasets, the first, annotated with 40 binary attributes such as hair color, gender, and age, and the other, with eight labels for facial expressions. In the experiments, StarGAN was compared with DIAT [Li

<sup>1</sup><https://www.kaggle.com/c/deepfake-detection-challenge>

<sup>2</sup><https://github.com/deepfakes/faceswap>

*et al.*, 2016], CycleGAN [Zhu *et al.*, 2017], IcGAN [Perarnau *et al.*, 2016], and other methods of transferring facial attributes. The results obtained in the CelebA dataset were evaluated using Amazon Mechanical Turk (AMT), where annotators selected the generated images with the highest degree of realism. The results obtained by StarGAN were the most chosen in both single-attribute and multi-attribute transfer tasks. For evaluation in the RaFD dataset, all experiments were trained on a ResNet-18 network using the same training and test sets. StarGAN produced the lowest classification error values, proving to be the model that generates the most realistic image translations with facial expressions among all models.

Nirkin *et al.* [2019] introduced Face Swapping GAN (FSGAN), a face-swapping model that retains poses and expressions from the original image. This model consists of a reenactment generator network to recreate the target face, a U-Net segmentation network to generate its segmentation mask, an inpainting network to estimate the possible occluded regions of the source face, and finally, a blending of the reenacted face into the target image using the segmentation network previously calculated. This model used IJB-C [Maze *et al.*, 2018] and LFW Parts Labels [Kae *et al.*, 2013] datasets to train the generator and segmentation networks, respectively. FSGAN was compared to deepfakes, Nirkin *et al.* [2018], and Face2Face methods during the experiments using the FaceForensics++ dataset. The results were evaluated by calculating the accuracy of head poses, the accuracy of facial expressions, the SSIM score, and face verification using Dlib. FSGAN achieved the best results in all metrics except for the face verification metric, where the values for the proposed model and deepfakes were the same.

Recently, Dhariwal and Nichol [2024] showed that diffusion models could achieve image sample quality superior to GANs. Adopting an UNet architecture for diffusion models, the authors proposed several changes, such as an increased number of attention heads; attention resolutions at 32x32, 16x16, and 8x8; the use of BigGAN [Brock *et al.*, 2019] residual block for upsampling and downsampling the activations; and rescaled residual connections. They also adopted a classifier guidance technique that trains a classifier on noisy images and then uses gradients to guide the diffusion sampling process toward an arbitrary class. The model has been trained on the ImageNet 128x128 dataset and evaluated through the FID metric, yielding 2.97 on ImageNet 128x128, 4.59 on ImageNet 256x256, and 7.72 on ImageNet 512x512. In this context, Rombach *et al.* [2022] introduced Stable Diffusion, a latent text-to-image diffusion model that was trained on 512x512 images from a subset of LAION-5B database [Schuhmann *et al.*, 2024]. This model uses a frozen CLIP ViT-L/14 text encoder to condition the model on text prompts and cross-attention layers into the model architecture. Stable Diffusion achieved state-of-the-art image inpainting and class-conditional image synthesis, being able to produce realistic images.

## 2.2 Methods for Facial Deepfake Detection

Afchar *et al.* [2018] introduced Meso-4 and MesoInception-4, two image classification-based networks. Both analyze

faces at a mesoscopic level and can detect fake features in videos without being compromised by image degradation generated by video compression. Experiments were conducted on two datasets. The first one, the deepfake dataset, consists of deepfake and real videos publicly available on the internet, while the other used a subset of Face2Face videos from the FaceForensics dataset. MesoInception-4 network achieved the best results in the deepfake dataset, with a classification score of 0.984 against 0.969 for Meso-4. In the Face2Face dataset, the results were the same, with both networks achieving a score of 0.953.

Tjon *et al.* [2021] proposed an architecture for deepfake detection that combines the EfficientNet B4 encoder pre-trained on ImageNet, and the Y-Net [Mehta *et al.*, 2018], which adds a classification branch at the end of the encoder to assign a real or fake label to the analyzed video frames, in addition to using a decoder trained to detect altered pixels existing in the images with the aid of segmentation masks.

EfficientNet family networks have become state-of-the-art in image recognition tasks, in the top 5 best solutions in the Kaggle DFDC [Dolhansky *et al.*, 2020b] competition. Pokroy and Egorov [2021] conducted a comparative study on the performance of different versions of EfficientNet, which vary according to the dimensions of the input data and the number of trainable parameters. They performed experiments on classification models trained for twenty epochs at DFDC and used Efficient baseline B0 to B7. They concluded that the networks with the most parameters only sometimes obtain the best results.

Recently, the transformer technique, initially proposed in natural language processing, has also been explored in computer vision tasks. Heo *et al.* [2021] presented an architecture that combines the Vision Transformer (ViT) and EfficientNet B7 pre-trained at DFDC. Using the global pooling operation, they merge the embeddings extracted by ViT and EfficientNet, passing them on to the transformer encoder. This method proved to have a superior performance in the deepfake detection task compared to state-of-the-art, with slightly better AUC and F1 scores.

Wang *et al.* [2022] proposed a transformer-based multi-scale architecture to identify regions synthesized by generative models. The model has two streams, one to capture fake features in the RGB domain and another to filter them in the frequency domain. The information obtained in both streams is combined through a cross-mode fusion block and passed to a fully connected layer and to a decoder that predicts the manipulated regions of the image through pixel difference masks. The authors also introduced SR-DF, a large-scale deepfake dataset built on FaceForensics++ videos. The proposed model was evaluated in the SR-DF, FaceForensics++, Celeb-DF [Li *et al.*, 2020b], and ForgeryNet datasets. The results achieved an AUC score of 99.92% on FaceForensics++, 95.5% on Celeb-DF, 86.7% on SR-DF, and 82.52% on ForgeryNet.

Coccomini *et al.* [2022] proposed two architectures based on CNN and vision transformer, Efficient ViT and Convolutional Cross ViT. The first one comprises a convolution module that uses a pre-trained EfficientNet B0 network and a vision transformer that classifies faces as real or fake through a CLS token. The second architecture is an adaptation of

Efficient ViT with two distinct branches, one to deal with the smaller features and the other to deal with the larger ones. The networks were trained from 220,444 faces extracted from the DFDC and FaceForensics++ datasets. The models were evaluated in the test set of both datasets and compared with other state-of-the-art models. The Convolutional Cross ViT network obtained the best results among the two proposed networks with an AUC of 0.51 and an F1-score of 88% in the DFDC dataset and an accuracy of 80% in the FaceForensics++ dataset.

Zhao *et al.* [2022] proposed a self-supervised transformer with a Contrastive Learning strategy to detect deepfake videos through features obtained from lip movement. Its architecture consists of two encoders, one for audio and another for video, both followed by an MLP projection head. The model was pre-trained on VoxCeleb2 and AV Speech datasets and fine-tuned on FaceForensics++. It was evaluated on the test set of the latter dataset at three compression levels, being compared to a supervised version of it and the state-of-the-art supervised model for lip reading tasks. Therefore, the proposed model reached results superior to the supervised one and close to the state-of-the-art, surpassing it in uncompressed videos with an ACC of 99.2% against 98.9% of the state-of-the-art.

In contrast to these works, the approach proposed by Gomes *et al.* [2023] uses self-supervised attention features as an input channel, along with RGB images, to differentiate between real images and deepfakes.

### 2.3 Fairness in Deepfake Detection

Trinh and Liu [2021] presented a thorough measure and analysis of the predictive performance of popular deepfake detectors on racially aware datasets balanced by gender and race. They trained MesoInception4 [Afchar *et al.*, 2018], Xception [Rossler *et al.*, 2019] and Face X-Ray [Li *et al.*, 2020a] on the FaceForensics++ dataset and cross-tested the models' generalizability with Google's DeepfakeDetection [Dufour and Gully, 2019], Celeb-DF, and DeeperForensics-1.0 [Jiang *et al.*, 2020]. Their findings point out important discoveries, such as significant disparities in predictive performances across races and a large bias representation in the commonly used FaceForensics++, composed mostly of Caucasian subjects with the majority of female Caucasian subjects. They also claim there is systematic discrimination towards female Asian subjects when detectors are trained with the Blended Images from Face X-rays.

Xu *et al.* [2024] also investigated biases regarding demographic and non-demographic attributes in public deepfake datasets and state-of-the-art deepfake detection models. First, They provide a massive annotation of five public datasets (Celeb-DF, DeepFakeDetection (DFD), FaceForensics++ (FF++), DeeperForensics-1.0 (DF-1.0), and Deepfake Detection Challenge Dataset (DFDC)) with 47 different attributes. They trained the three models EfficientNetB0 [Tan and Le, 2019], Xception, and Capsule-Forensics-v2 [Nguyen *et al.*, 2019]. Their analysis indicates that the datasets used lack diversity and that the deepfake detection models demonstrate strong bias issues for many demographic and non-demographic attributes. Due to this,

they claim that the biased performance may lead to significant societal fairness and security issues depending on the use case. Furthermore, imbalanced attributes within these datasets could exacerbate generalization problems across various attributes in contemporary Deepfake detection algorithms.

These other works address performance differences in many models with different architectures, yet they do not encompass those with self-supervised features. This highlights the need for further study into how this technique affects the demographic imbalances seen in other works.

## 3 Method

Figure 1 illustrates the architecture of our proposal for deepfake detection. Given an RGB image, a Face Detector extracts a patch  $x \in \mathbb{R}^{(H \times W \times 3)}$  of a face. Next, the Self-supervised Facial Model generate self-supervised features  $x_{am} \in \mathbb{R}^{(H \times W \times 1)}$  from the given patch. Then, both  $x$  and  $x_{am}$  are concatenated, producing a tensor  $x_c \in \mathbb{R}^{(H \times W \times 4)}$  which is finally used to feed a discriminator that produces scores for real and fake categories.

In the remainder of this section, we detail the modules composing our proposal's architecture. We introduce the self-supervised facial feature extractor in Section 3.1. Next, we describe the CNN-based models used to classify deepfakes in Section 3.2.

### 3.1 Self-Supervised Facial Feature Extractor

We used the DINOv2 [Oquab *et al.*, 2024] model to extract the self-attention activation maps from images. The DINO family architecture comprises two neural networks, the student and the teacher, which share the same ViT architecture but different parameters. The ViT architecture used in DINO takes a grid of non-overlapping contiguous image patches as input and then projects an attention head at its output. Therefore, a multi-crop strategy generates different views of the input image, which are then passed to both networks, generating probability distributions by normalizing the networks' output with a softmax function. Besides that, part of DINO's good results is due to using a momentum encoder in the teacher network. It must also center and sharpen operations applied to its outputs to avoid collapse.

DINOv2 significantly outperforms its first version, DINO [Caron *et al.*, 2021], in video segmentation quality through three main improvements: a vastly more significant and diverse training dataset called LVD-142M with 142 million images, enhanced training algorithms and implementation techniques using PyTorch<sup>3</sup> and xFormers<sup>4</sup> for better stability and efficiency, and an advanced knowledge distillation process for compressing large models into smaller ones without substantial accuracy loss. These enhancements contribute to DINOv2's superior understanding, segmentation capabilities, and performance across various tasks, maintaining high efficiency even with smaller model sizes.

<sup>3</sup><https://pytorch.org/get-started/pytorch-2.0/>

<sup>4</sup><https://github.com/facebookresearch/xformers>

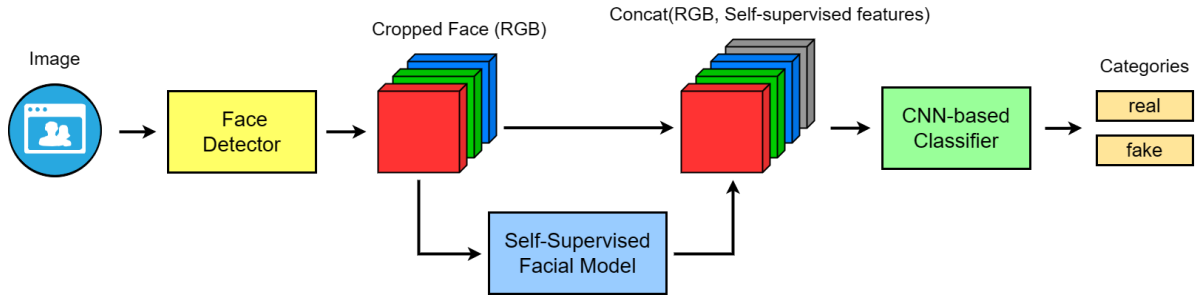


Figure 1. Method for deepfake classification. Source: Gomes *et al.* [2023].

The authors provided on their GitHub Repository<sup>5</sup> the weights of the pre-trained models both in the ViT-Base architecture, with 86M parameters and in the ViT-Small, with 21M parameters. We opted to use the pre-trained model ViT-S/14 for extracting the self-attention maps from our dataset.

Gomes *et al.* [2023] employed transfer learning from the DINO’s pre-trained model to generate three different attention heads with self-supervised facial features for each facial image of our dataset. Each resulting attention head was taken individually as the fourth channel of the model’s inputs. Figure 2 illustrates the self-attention activation maps extracted by DINO. These original results indicated that attention head 1 outperformed others in most of the tests. Therefore, we limited the scope of this reevaluation to focus solely on attention head 1.

## 3.2 CNN-based Classifier

This section details the different CNNs initially tested as possible backbones for the CNN-based deepfake classifier. In this work, we focused on the EfficientNet B4 and Xception due to their higher performances in our initial tests, their prevalence among other works in deepfake detection [Bonetini *et al.*, 2021; Rossler *et al.*, 2019], and their presence in other demographic analyses of deepfake detection [Xu *et al.*, 2024; Trinh and Liu, 2021].

### 3.2.1 EfficientNet B4

Introduced by Tan and Le [2019], EfficientNet B4 is part of the EfficientNet model family, which seeks a balance between performance and computational cost, having different versions (B0-B7) that vary the size of the input data and the number of trainable parameters. One of the main features of EfficientNet B4 is the use of mixed depthwise convolution blocks, which combine depthwise separable convolutions with standard convolutions. This combination significantly reduces the computational cost without compromising the quality of visual feature representation.

EfficientNet B4 was designed through an automated optimization process that simultaneously adjusts the models’ depth, width, and resolution scale, allowing them to reach maximum performance with relatively few parameters. Additionally, the EfficientNet B4 network has shown one of the best performances in Pokroy and Egorov [2021]. Therefore, we opted for EfficientNet B4 to balance performance and training speed.

### 3.2.2 Xception

Xception [Chollet, 2017] consists of a deep convolutional neural network architecture developed by Google researchers. In this network, Inception modules have been replaced with depthwise separable convolutions, an intermediate point between them and regular convolution. The depthwise convolution is applied on a one-by-one channel instead of that application, which uses it for all channels.

The approach used by Xception allows a significant reduction in computational cost, as operations on separate channels can be performed in parallel and more efficiently in terms of techniques. This technique was innovative for convolutional neural networks as it leverages the advantages of depthwise separable convolutions and residual connections to improve efficiency and the ability to represent visual features while maintaining learning capacity and feature representation despite the reduced computational cost.

## 4 Experiments

In this section, we go over the steps we took to expand on the results from Gomes *et al.* [2023] as well as the new data used to further validate the performance of the original proposal. The first subsection covers the new dataset we used to demonstrate the consistency of our original findings. The second subsection deals with the demographic data in the dataset and elaborates on how we divided it to determine the patterns that emerged, as seen in Figure 3.

### 4.1 Dataset

Similarly to Gomes *et al.* [2023], we used the selection of the Deepfake Detection Challenge Dataset (DFDC) [Dolhansky *et al.*, 2020a] due to its diversity of lighting conditions, resolutions, and people of different genders, ages, and skin colors, with the demographic diversity being of particular importance to this study. We expanded our data while still using the same dataset by using the over 124,000 videos in the dataset and extracting more frames from each video, with an average of around 10 frames per video, meaning that the same face would appear in multiple images on the final dataset, but always confined to the same subset (training, validation or testing). We used the OpenCV<sup>6</sup> package to extract the frames from each video to preprocess the dataset. In addition, we used the Multitask Cascaded Convolutional

<sup>5</sup><https://github.com/facebookresearch/dinov2>

<sup>6</sup><https://opencv.org/>



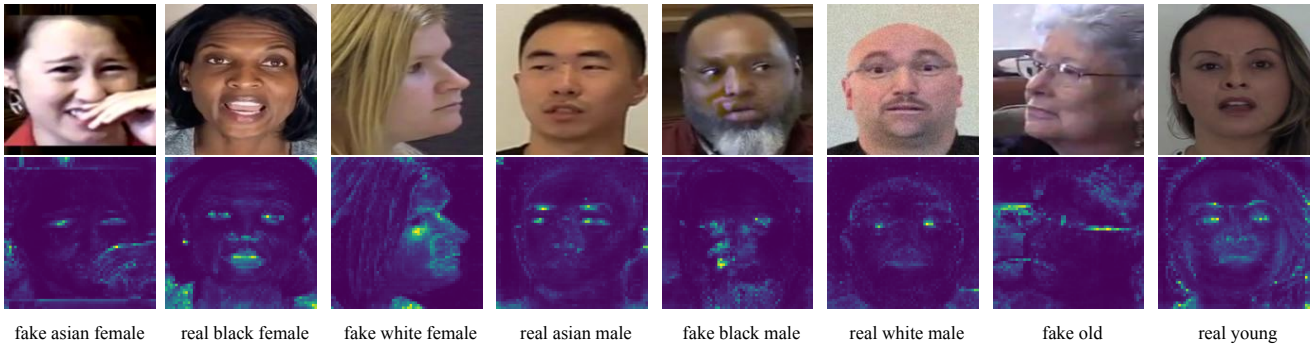


Figure 2. Images from different demographic groups in the dataset and their respective attention heads.

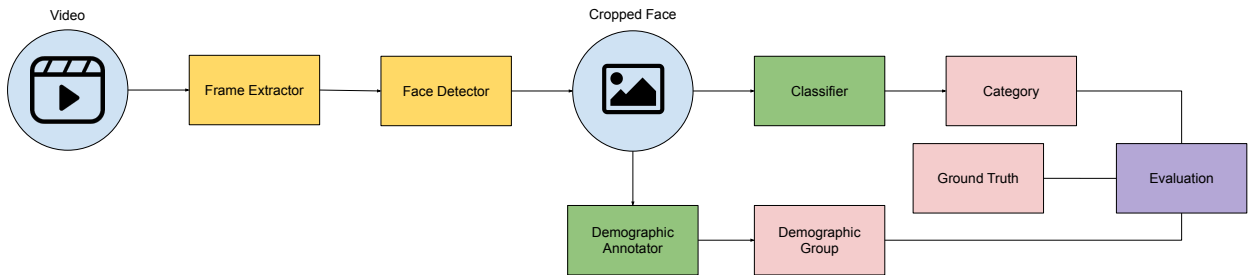


Figure 3. Dataflow in validation experiments.

Network (MTCNN) face detector [Zhang *et al.*, 2016] with thresholds (0.95, 0.95, 0.95) and margin 40 to identify and crop faces within a video frame. While a full performance evaluation of the face detector fell outside the scope of this work, we observed minimal to no issues while sampling its outputs. We then obtained images in *JPG* format of the facial regions for each extracted frame, totaling 1,186,737 images in the training set and 44,316 in the validation set.

Unbalanced deepfake detection datasets can produce biased models [Xu *et al.*, 2024]. Therefore, it is important to understand the demographic distribution of the dataset. We used CLIP [Radford *et al.*, 2021] to annotate our dataset regarding the two main demographic attributes of apparent gender and ethnicity [Trinh and Liu, 2021] and the attribute of age cited by Xu *et al.* [2024]. The distribution of each demographic attribute for our validation set is presented in Table 1. The dataset was mostly balanced in terms of apparent gender and age, with ethnicity remaining the main point of divergence. It is important to keep these disparities in mind when discussing the results. While demographic data was used in our final analysis, it was not part of the model’s inputs, which consisted only of a frame and its attention head.

## 4.2 Setup

In our experiments, we trained two different models for each architecture presented in Section 3.2: a baseline model trained only with 3-channel inputs, corresponding to RGB facial images, and a model using the attention map presented in Section 3.1 as the fourth channel. Thus, it was possible to verify the gain of using the self-supervised features compared to the baseline models, as seen in Gomes *et al.* [2023].

We used the default input sizes of each architecture de-

Table 1. Demographic distribution of validation set.

Attribute	Frames	
	Amount	Percentage (%)
Overall	44,316	100.00
Female	22,978	51.85
Male	21,338	48.15
Old	22,716	51.26
Young	21,600	48.74
Asian	11,454	25.85
Black	14,674	33.11
White	18,188	41.04

scribed earlier for training and evaluation steps, an Adam optimizer with a learning rate of  $1e-4$ , and a categorical cross-entropy loss function.

The experiments were run on a system with 48GB of system memory, 850GB of storage, and a Nvidia RTX 2080 11GB.

## 4.3 Validation Results

The goal of this validation was to ensure that the proposal from Gomes *et al.* [2023] would scale properly into the new dataset, maintaining its original level of performance. For this, we evaluated our proposed models based on the EfficientNet B4 [Tan and Le, 2019] and Xception [Chollet, 2017] on a new validation set, using the AUC and F1-Score (weighted) metrics. The results confirmed the increase in performance at levels similar to those previously obtained when compared to the base models. Table 2 shows the results for this validation, where, as expected, the Xception-

based model outperformed the one based on the EfficientNet B4, mirroring the previous results from [Gomes *et al.*, 2023].

## 5 Demographic Analysis

Following growing concerns about demographic bias in deepfake detection [Xu *et al.*, 2024; Trinh and Liu, 2021], we aim to show how our proposed approach deals with different demographic attributes. This section explores the results of our demographic experiments and discusses how they relate to other studies in the area.

For the specific purpose of this analysis, we focused on the combination of Xception and attention head 1, as they achieved our best overall results with the full dataset. Table 3 shows how its performance varied across different combinations of our chosen demographic attributes. Additionally, Figure 5 shows a few examples of instances where the model was mistaken in its classification.

### 5.1 Overall Results

The first thing to note is that our approach proved consistent when considering each attribute in isolation. Considering AUC, every group remained within 0.51 percentage points of the overall results, with the largest difference being 1.01% between the classes Male and Female. This results in a range of values between 91.60% and 92.61% and a standard deviation of 0.34. Apparent gender proved to be the only statistically significant variable when considering each attribute in isolation, with the results for age and ethnicity being inconclusive.

This small variance speaks to the consistency of our approach, but looking at each attribute in isolation does not paint a full picture. It is also important to consider different combinations of attributes, as they can more closely represent one specific group of people. In this aspect, as expected, we had more varied results, ranging from 88.75% to 93.99% with a standard deviation of 1.25. Most combinations of two attributes proved to have a statistically significant impact on the results with a 95% level of confidence, with the exceptions of female + asian, female + old, old + white, and young + white.

Notably, two groups underperformed both the overall results and the other below-average groups, those being *Black Male* and *White Female*. It is worth mentioning that both ethnicities when combined with the two apparent genders, had slightly below-average results. Therefore, the highly negative results of these groups, along with the positive ones from *White Male* and *Black Female*, show that both ethnicities had detection issues that were more localized to a specific apparent gender instead of applying to all instances of that ethnicity.

In terms of age, we saw that the attribute alone had little effect on the results, instead impacting how other attributes affect performance. Age had inverse effects on each apparent gender, with *Young* having a positive effect for *Male* and negative for *Female*, with the opposite being true for *Old*. As for their impact on ethnicity, *Old* had a positive effect for *Black*, while *Young* had the same for *Asian* and *White*. *Old*

was the more consistent of the two, with a standard deviation of 0.93 across the different apparent genders and ethnicities, while *Young* had 1.38. This indicates that age is the most consistent attribute, reinforced by the fact that its biggest outlier was the case of *Young Male*, which was 1.47 standard deviations above the overall results, contrasted to *White Female*, who was 1.88 standard deviations below the overall results. This contrasts with apparent gender since *Female* and *Male* have standard deviations of 1.55 and 1.52, respectively, across the different apparent ages and ethnicities, with both being higher than the values presented for both age groups.

Back to the topic of ethnicity, it is worth noting that *Asian*, besides having the overall best results, was also the most consistent class with a standard deviation of 1.18 across the different apparent genders and ages, compared to 1.51 for *Black* and 1.77 for *White*. The high performance of *Asian*, the least represented ethnicity, calls into question the relationship between performance in a given group and its support. However, as seen in Figure 4, we found no correlation between these factors.

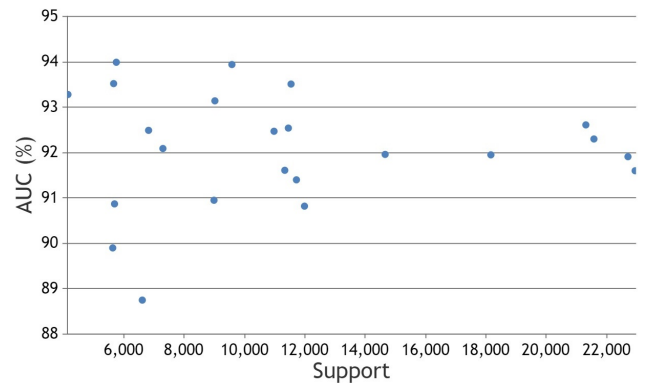


Figure 4. Relationship between AUC and support. Each point represents a different demographic group, shown in Table 3.

### 5.2 Class Specific Results

The results from the last subsection, while already giving a pretty clear indication of the consistency of our approach, still do not encompass the whole story. Different applications of deepfake detection have different priorities regarding how much they tolerate errors for real and fake submissions. For instance, social media platforms do not want genuine users to have their posts frequently marked as fake [Trinh and Liu, 2021]. The metrics used in Table 3, by their definition, do not show the details of how the approach performs depending on the target label, making it important for us to explicitly examine precision and recall to determine how each group is affected in different scenarios. Table 4 shows how the demographic attributes affect precision and recall. In our experiments, detecting an entry as fake was considered a positive detection, while detecting it as real was considered a negative one, so the lower the precision, the more likely it is for real images to be considered fake, and the lower the recall, the more likely it is for fake images to be allowed as real.

The first thing to note is that, similarly to the first table, each attribute in isolation was, in terms of precision, close

**Table 2.** Overall results.

Model	AUC (%)	Diff. (%)	F1-Score (%)	Diff. (%)
EfficientNet B4 Baseline	90.78	---	83.32	---
EfficientNet B4 + Head 1	90.92	+0.14	83.39	+0.07
Xception Baseline	91.91	---	84.36	---
Xception + Head 1	92.10	+0.19	84.68	+0.32

**Table 3.** Xception performance across different demographics.

Attribute	Frames	AUC (%)	Diff. (%)	F1-Score (%)	Diff. (%)
Overall	44,316	92.10	---	84.68	---
Female	22,978	91.60	-0.50	84.28	-0.40
Male	21,338	92.61	+0.51	85.09	+0.41
Old	22,716	91.91	-0.19	84.26	-0.42
Young	21,600	92.30	+0.20	85.01	+0.33
Asian	11,454	92.54	+0.44	85.64	+0.96
Black	14,674	91.96	-0.14	84.15	-0.53
White	18,188	91.95	-0.15	84.44	-0.24
Female + Asian	7,314	92.09	-0.01	85.21	+0.53
Female + Black	9,040	93.14	+1.04	85.96	+1.28
Female + White	6,624	88.75	-3.35	81.05	-3.63
Female + Old	10,977	92.47	+0.37	85.26	+0.58
Female + Young	12,001	90.82	-1.28	83.22	-1.46
Male + Asian	4,144	93.28	+1.18	86.44	+1.76
Male + Black	5,634	89.90	-2.20	81.34	-3.34
Male + White	11,564	93.51	+1.41	86.38	+1.70
Male + Old	11,739	91.40	-0.70	83.33	-1.35
Male + Young	9,599	93.94	+1.84	87.20	+2.52
Old + Asian	5,702	90.87	-1.23	83.06	-1.62
Old + Black	5,672	93.52	+1.42	86.25	+1.57
Old + White	11,342	91.61	-0.49	86.80	+2.12
Young + Asian	5,752	93.99	+1.89	88.09	+3.41
Young + Black	9,002	90.95	-1.15	82.70	-1.98
Young + White	6,846	92.49	+0.39	85.41	+0.73

to the overall results with a standard deviation of 1.31. The biggest outliers were *Young* and *Black*, with precision 2.39% lower and 1.77% higher than the overall results, respectively. However, the same consistency was not seen in the context of recall, which had a standard deviation of 4.13. The worst results were also from *Young* and *Black*, with recall 5.91% and 5.45% lower than the overall results, respectively. In the specific case of *Black* demographic, the low recall combined with a slightly above-average precision indicates that these images were more likely to be classified as real rather than as fake, with this group being the only standalone attribute to show this pattern. The opposite pattern emerged in the case of *Asian*, with a recall much higher than precision, indicating that these images were more likely to be flagged as fake.

When considering different combinations of attributes, we again saw more varied results. Precision ranged from 76.57% to 90.63%, with a standard deviation of 3.59, and recall was again the less consistent of the two, ranging from 71.94% to 92.28% and a standard deviation of 5.99, indicat-

ing that real images are on average less affected by demographic attributes.

In terms of precision, some of the worst results we saw were for *Black Male* and *White Female*, with 76.57% and 80.72%, respectively, mirroring how these two groups also had the worst results in terms of AUC. Their similarly low recall shows performance issues that were not made clear, at least not to their true extent, by aggregate metrics such as AUC and F-Score.

*Asian Female* also saw particularly low precision (78.69%), but their high recall (87.19%) indicates that their images were considerably more likely to be classified as fake. A similar but less intense pattern was seen in *Young Asian* with 80.38% precision and 88.33% recall, which isn't entirely surprising considering the overlap between the two groups.

In terms of recall, the two groups with the worst performances were, by far, *Young Black* and *Young Female* (71.94% and 73.39%, respectively). This does not come as a



**Table 4.** Xception precision and recall across different demographics.

Attribute	Frames	Precision (%)	Diff. (%)	Recall (%)	Diff. (%)
Overall	44,316	83.78	---	84.12	---
Female	22,978	84.13	+0.35	81.84	-2.28
Male	21,338	83.44	-0.34	86.48	+2.36
Old	22,716	85.18	+1.40	87.82	+3.70
Young	21,600	81.39	-2.39	78.21	-5.91
Asian	11,454	83.08	-0.70	89.63	+5.51
Black	14,674	85.55	+1.77	78.67	-5.45
White	18,188	83.04	-0.74	84.67	+0.55
Female + Asian	7,314	78.69	-5.09	87.19	+3.07
Female + Black	9,040	90.63	+6.85	79.88	-4.24
Female + White	6,624	80.72	-3.06	79.65	-4.47
Female + Old	10,977	86.19	+2.41	88.04	+3.92
Female + Young	12,001	80.97	-2.81	73.39	-10.73
Male + Asian	4,144	88.12	+4.34	92.28	+8.16
Male + Black	5,634	76.57	-7.21	76.26	-7.86
Male + White	11,564	83.77	-0.01	87.75	+3.63
Male + Old	11,739	84.27	+0.49	87.61	+3.49
Male + Young	9,599	81.86	-1.92	84.33	+0.21
Old + Asian	5,702	84.66	+0.88	90.37	+6.25
Old + Black	5,672	87.78	+4.00	86.63	+2.51
Old + White	11,342	84.27	+0.49	86.90	+2.78
Young + Asian	5,752	80.38	-3.40	88.33	+4.21
Young + Black	9,002	83.40	-0.38	71.94	-12.18
Young + White	6,846	79.73	-4.05	78.94	-5.18

surprise, considering how all three attributes had lower-than-average recall. Both groups also had lower-than-average precision, but just slightly so in the case of *Young Black*. It is also worth noting that both *Old Female* and *Old Black* had higher-than-average recall, indicating this issue might be related to their combination with *Young* age.

When considering standalone attributes, apparent gender was the most consistent both in terms of precision and recall, with standard deviations of 0.34 and 2.32 respectively. Age was the most inconsistent for both values, with standard deviations of 1.89 and 4.81, respectively, with the particularly high recall standard deviation helping reinforce the results mentioned in the previous paragraph.

When it comes to ethnicity, *White* had the most consistent precision with a standard deviation of 1.94, while Asian had the most consistent recall with a standard deviation of 1.95. This shows that the ability to deal with real images from *White* remained mostly consistent but also mostly below average. Meanwhile, the ability to deal with fake images from *Asian* was also fairly consistent and above average. *Black*, on the other hand, had the most varied results for both metrics, being highly impacted by apparent gender and age, resulting in standard deviations of 5.30 and 5.38.

Lastly, Figure 5 shows a few examples of incorrectly classified images. Some of these contain clear visual artifacts or noise, which can disrupt the model and lead to misclassifications. Others represent more subtle patterns, such as slight

variations in facial expressions or lighting conditions, which might be challenging for the model to accurately distinguish. These indicate a need for further refinement, possibly using additional training data or feature extraction techniques, to improve robustness against these complex nuances.



fake faces classified as real      real faces classified as fake  
**Figure 5.** Some examples of images misclassified by the model.

## 6 Conclusion

With the ever-increasing generation of realistic synthetic media, there is an ongoing challenge to distinguish between genuine and manipulated content, and deepfake detection technologies emerge as potential tools to address the adverse effects of the possible spreading of harmful information, including fake news and incitements to violence.

By extending Gomes *et al.* [2023], the present work explores how demographic classes of attributes behave within a classifier that applies self-supervised features generated by a foundation model in the task of realistic facial deepfake detection. In summary, the present study extends upon the findings of that work by (i) making use of an expanded dataset to validate the original results further, (ii) applying a new upgraded version of DINO (DINOv2 [Oquab *et al.*, 2024]), and (iii) investigating the new following research question: *How do demographic classes of attributes behave within a self-supervised strategy aimed at classifying realistic facial deepfakes?*

The results indicate that performance remains consistent across most demographic classes, with minimal variation observed when considering the difference between specific and overall performances. It also shows that performance can vary between different classes (real or fake) within the same demographic group, potentially affecting the user experience of certain groups, depending on the specific use case of the deepfake detection model. In addition, our results reveal that an imbalance in dataset distribution can lead to biased models. However, their existence alone does not always explain such results, as we saw no correlation between support for a demographic group and our model's performance in that group.

Looking ahead, several potential opportunities for future research deserve attention. One promising direction involves applying the findings from these analyses to improve the classifier's performance by targeting specific characteristics within the demographic classes that exhibited less satisfactory performances. Conducting new experiments would also be interesting, particularly those utilizing alternative datasets or new detection approaches, particularly ones employing ensemble techniques. This would allow us to confirm whether the observed impacts on performance are consistent across different perspectives or are particular to one of them. Besides that, as deepfake generation is still evolving, there is a constant need to explore newer models, particularly foundational ones such as Google's Gemini or Microsoft's LLaVa, to enhance detection performance further.

## Declarations

### Funding

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-22-1-0475, and also financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES)

## Authors' Contributions

Yan Gurevitz: Conceptualization, Methodology, Investigation, Software, Validation, Writing - original draft. Bruno Gomes: Conceptualization. Daniel Moraes, José Boaro, Antonio Busson, Julio Duarte, and Sérgio Colcher: Methodology, Writing - original draft, Writing – review & editing. All authors read and approved the final manuscript.

## Competing interests

The authors declare that they have no competing interests.

## Availability of data and materials

The datasets generated or analyzed during the current study are available at <https://www.kaggle.com/competitions/deepfake-detection-challenge/data>.

## References

- Afchar, D., Nozick, V., Yamagishi, J., and Echizen, I. (2018). Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE. DOI: <https://doi.org/10.1109/WIFS.2018.8630761>.
- Almond Solutions (2021). Why do people post on social media. <https://www.almondsolutions.com/blog/why-do-people-post-on-social-media>. Accessed: 09 July 2024.
- Beaumont-Thomas, B. (2024). Taylor swift deepfake pornography sparks renewed calls for us legislation. <https://www.theguardian.com/music/2024/jan/26/taylor-swift-deepfake-pornography-sparks-renewed-calls-for-us-legislation>. Accessed: 09 July 2024.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., *et al.* (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. DOI: <https://doi.org/10.48550/arXiv.2108.07258>.
- Bonettini, N., Cannas, E. D., Mandelli, S., Bondi, L., Bestagini, P., and Tubaro, S. (2021). Video face manipulation detection through ensemble of cnns. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5012–5019. DOI: <https://doi.org/10.1109/ICPR48806.2021.9412711>.
- Brock, A., Donahue, J., and Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. DOI: <https://doi.org/10.48550/arXiv.1809.11096>.
- Bulat, A. and Tzimiropoulos, G. (2017). How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Com-*

- puter Vision, pages 1021–1030. DOI: <https://doi.org/10.1109/ICCV.2017.116>.
- Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660. DOI: <https://doi.org/10.1109/ICCV48922.2021.00951>.
- Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). 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. DOI: <https://doi.org/10.1109/CVPR.2018.00916>.
- Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258. DOI: <https://doi.org/10.1109/CVPR.2017.195>.
- Coccomini, D. A., Messina, N., Gennaro, C., and Falchi, F. (2022). Combining efficientnet and vision transformers for video deepfake detection. In Sclaroff, S., Distanto, C., Leo, M., Farinella, G. M., and Tombari, F., editors, *Image Analysis and Processing – ICIAP 2022*, pages 219–229, Cham. Springer International Publishing. DOI: [https://doi.org/10.1007/978-3-031-06433-3\\_19](https://doi.org/10.1007/978-3-031-06433-3_19).
- Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., and Verdoliva, L. (2023). 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. DOI: <https://doi.org/10.1109/ICASSP49357.2023.10095167>.
- Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65. DOI: <https://doi.org/10.1109/MSP.2017.2765202>.
- Dhariwal, P. and Nichol, A. (2024). Diffusion models beat GANs on image synthesis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems, NIPS '21*, Red Hook, NY, USA. Curran Associates Inc.. DOI: <https://dl.acm.org/doi/10.5555/3540261.3540933>.
- Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C. C. (2020a). The deepfake detection challenge dataset.
- Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C. C. (2020b). The deepfake detection challenge (DFDC) dataset. <https://doi.org/10.48550/arXiv.2006.07397>. Accessed: 09 July 2024.
- Dufour, N. and Gully, A. (2019). Contributing data to deepfake detection research. <https://blog.research.google/2019/09/contributing-data-to-deepfake-detection.html?m=1>. Accessed: 09 July 2024.
- Feng, Y., Wu, F., Shao, X., Wang, Y., and Zhou, X. (2018). Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551. DOI: [https://doi.org/10.1007/978-3-030-01264-9\\_33](https://doi.org/10.1007/978-3-030-01264-9_33).
- Gomes, B. R., Busson, A. J. G., Boaro, J., and Colcher, S. (2023). Realistic facial deep fakes detection through self-supervised features generated by a self-distilled vision transformer. In *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web, WebMedia '23*, page 177–183, New York, NY, USA. Association for Computing Machinery. DOI: <https://doi.org/10.1145/3617023.3617047>.
- Heo, Y.-J., Choi, Y.-J., Lee, Y.-W., and Kim, B.-G. (2021). Deepfake detection scheme based on vision transformer and distillation. *arXiv preprint arXiv:2104.01353*. DOI: <https://doi.org/10.48550/arXiv.2104.01353>.
- Iglovikov, V. and Shvets, A. (2018). Ternaunet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746*. DOI: <https://doi.org/10.48550/arXiv.1801.05746>.
- Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134. DOI: <https://doi.org/10.1109/CVPR.2017.632>.
- Jiang, L., Li, R., Wu, W., Qian, C., and Loy, C. C. (2020). Deepforensics-1.0: A large-scale dataset for real-world face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2886–28958. DOI: <https://doi.org/10.1109/CVPR42600.2020.00296>.
- Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer. DOI: [https://doi.org/10.1007/978-3-319-46475-6\\_43](https://doi.org/10.1007/978-3-319-46475-6_43).
- Kae, A., Sohn, K., Lee, H., and Learned-Miller, E. (2013). Augmenting CRFs with boltzmann machine shape priors for image labeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2019–2026. DOI: <https://doi.org/10.1109/CVPR.2013.263>.
- Khalid, H. and Woo, S. S. (2020). Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 656–657. DOI: <https://doi.org/10.1109/CVPRW50498.2020.00336>.
- Kiefer, B. (2023). This brand's social experiment uses ai to expose the dark side of 'sharenting'. <https://www.adweek.com/brand-marketing/this-brands-social-experiment-uses-ai-to-expose-the-dark-side-of-sharenting/>. Accessed: 09 July 2024.
- King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758. DOI: <https://dl.acm.org/doi/10.5555/1577069.1755843>.
- Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., and Van Knippenberg, A. (2010). Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388. DOI: <https://doi.org/10.1080/02643758.2010.508686>.

- [//doi.org/10.1080/02699930903485076](https://doi.org/10.1080/02699930903485076).
- Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., and Guo, B. (2020a). Face x-ray for more general face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5000–5009. DOI: <https://doi.org/10.1109/CVPR42600.2020.00505>.
- Li, M., Zuo, W., and Zhang, D. (2016). Deep identity-aware transfer of facial attributes. *arXiv preprint arXiv:1610.05586*. DOI: <https://doi.org/10.48550/arXiv.1610.05586>.
- Li, Y., Sun, P., Qi, H., and Lyu, S. (2022). Toward the creation and obstruction of deepfakes. In *Handbook of Digital Face Manipulation and Detection*, pages 71–96. Springer, Cham. DOI: [https://doi.org/10.1007/978-3-030-87664-7\\_4](https://doi.org/10.1007/978-3-030-87664-7_4).
- Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. (2020b). CelebDF: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3204–3213. DOI: <https://doi.org/10.1109/CVPR42600.2020.00327>.
- Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*. DOI: <https://doi.org/10.1109/ICCV.2015.425>.
- Maze, B., Adams, J., Duncan, J. A., Kalka, N., Miller, T., Otto, C., Jain, A. K., Niggel, W. T., Anderson, J., Cheney, J., et al. (2018). IARPA janus benchmark - c: Face dataset and protocol. In *2018 international conference on biometrics (ICB)*, pages 158–165. IEEE. DOI: <https://doi.org/10.1109/ICB2018.2018.00033>.
- Mehta, S., Mercan, E., Bartlett, J., Weaver, D., Elmore, J. G., and Shapiro, L. (2018). Y-net: joint segmentation and classification for diagnosis of breast biopsy images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 893–901. Springer. DOI: [https://doi.org/10.1007/978-3-030-00934-2\\_99](https://doi.org/10.1007/978-3-030-00934-2_99).
- Nguyen, H. H., Yamagishi, J., and Echizen, I. (2019). Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311. DOI: <https://doi.org/10.1109/ICASSP.2019.8682602>.
- Nirkin, Y., Keller, Y., and Hassner, T. (2019). FSGAN: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193. DOI: <https://doi.org/10.1109/ICCV.2019.00728>.
- Nirkin, Y., Masi, I., Tuan, A. T., Hassner, T., and Medioni, G. (2018). On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 98–105. IEEE. DOI: <https://doi.org/10.1109/FG.2018.00024>.
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2024). DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*. DOI: <https://doi.org/10.48550/arXiv.2304.07193>.
- Perarnau, G., Van De Weijer, J., Raducanu, B., and Álvarez, J. M. (2016). Invertible conditional GANs for image editing. *arXiv preprint arXiv:1611.06355*. DOI: <https://doi.org/10.48550/arXiv.1611.06355>.
- Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C. S., RP, L., Jiang, J., et al. (2023). Deepfacelab: Integrated, flexible and extensible face-swapping framework. *Pattern Recogn.*, 141(C). DOI: <https://doi.org/10.1016/j.patcog.2023.109628>.
- Pokroy, A. A. and Egorov, A. D. (2021). Efficient-nets for deepfake detection: Comparison of pretrained models. In *2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, pages 598–600. IEEE. DOI: <https://doi.org/10.1109/ElConRus51938.2021.9396092>.
- Radford, A., Kim, J. W., Chris Hallacy, A. R., Gabriel Goh, S. A., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR. DOI: <https://doi.org/10.48550/arXiv.2103.00020>.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685. DOI: <https://doi.org/10.1109/CVPR52688.2022.01042>.
- Rosler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11. DOI: <https://doi.org/10.1109/ICCV.2019.00009>.
- Schmunk, R. (2024). Explicit fake images of Taylor Swift prove laws haven't kept pace with tech, experts say. <https://www.cbc.ca/news/canada/taylor-swift-ai-images-highlight-need-for-better-legislation-1.7096094>. Accessed: 09 July 2024.
- Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. (2024). LAION-5B: An open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*. DOI: <https://dl.acm.org/doi/10.5555/3600270.3602103>.
- Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR. DOI: <https://doi.org/10.48550/arXiv.1905.11946>.
- Tjon, E., Moh, M., and Moh, T.-S. (2021). Eff-yonet: A dual task network for deepfake detection and seg-

- mentation. In *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pages 1–8. IEEE. DOI: <https://doi.org/10.1109/IMCOM51814.2021.9377373>.
- Trinh, L. and Liu, Y. (2021). An examination of fairness of ai models for deepfake detection. In Zhou, Z.-H., editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 567–574. International Joint Conferences on Artificial Intelligence Organization. Main Track. DOI: <https://doi.org/10.24963/ijcai.2021/79>.
- Wang, J., Wu, Z., Chen, J., and Jiang, Y.-G. (2022). M2TR: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, page 615–623. DOI: <https://doi.org/10.1145/3512527.3531415>.
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612. DOI: <https://doi.org/10.1109/TIP.2003.819861>.
- Xu, Y., Terhörst, P., Raja, K., and Pedersen, M. (2024). Analyzing fairness in deepfake detection with massively annotated databases. *IEEE Transactions on Technology and Society*, 5(1):93–106. DOI: <https://doi.org/10.1109/TTS.2024.3365421>.
- Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503. DOI: <https://doi.org/10.1109/LSP.2016.2603342>.
- Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., and Li, S. Z. (2017). S3FD: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201. DOI: <https://doi.org/10.1109/ICCV.2017.30>.
- Zhao, H., Zhou, W., Chen, D., Zhang, W., and Yu, N. (2022). Self-supervised transformer for deepfake detection. *arXiv preprint arXiv:2203.01265*. DOI: <https://doi.org/10.48550/arXiv.2203.01265>.
- Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2242–2251. DOI: <https://doi.org/10.1109/ICCV.2017.244>.