Performance Comparison of Synthetic Face Databases using the Xception Model

Evaluating protection against slander

Filip Dobrev [1]
Supervisor(s): Anna Lukina[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Filip Dobrev
Final project course: CSE3000 Research Project
Thesis committee: Anna Lukina, Petr Kellnhofer

Abstract

The rise of AI-generated images presents significant challenges in distinguishing between real and fake visuals. Such fake content can disseminate false information about someone or create false identities for fraud. This study evaluates the effectiveness of the Xception model in detecting AI-generated faces, a crucial task to mitigate the misuse of facial manipulation technology. By analyzing various datasets, including iFakeFaceDB, Diverse Fake Face Dataset (DFFD), CelebA, and CASIAWebFace, we assess the model's performance in real-world scenarios. Our findings highlight the strengths of the Xception model at recognising real and synthetic images. It achieves an accuracy upwards of 97.11% on DFFD, which includes both real and synthetic images, and 96.87%-98.07% on CASIA and CelebA. Furthermore, it achieves an accuracy of 73.15% on a different synthetic facial dataset - iFakeFaceDB. This work examines the Xception model's capabilities and underscores the need for comprehensive detection methods to safeguard against the potential harms of synthetic media.

1 Introduction

The rise of AI-generated images is a growing concern in our society [1]. Generative Adversarial Networks (GANs) can create lifelike pictures that cannot be distinguished from real ones. In a recent study Miller et. al. [1] coined the term "AI hyperrealism" meaning that an AI is able to create faces indistinguishable from humans. In their results in Figure 1, we can observe that four out of the top five faces most often regarded as human are artificially created. Face2Face [2] is another technology that can manipulate a person's photos, by transferring their face to another person. Such images can depict people committing crimes, spread slander, and influence political campaigns. This technology, commonly known as deepfake, is often used to create explicit content of celebrities, such as in a recent incident with Taylor Swift [3]. This incident proves the influence such images can have on a person's life. What is alarming is the accessibility of this technology to the public. Anyone can easily download a text-to-image model, as many projects are open source[1]. The processing power requirements are low as an Nvidia RTX 3060 GPU[2] is more than capable of running the model[1]. This gives a lot of power to malicious people to abuse. Therefore we need urgent action for proactive measures to fight these emerging issues.

The ability to quickly detect such images in social media is essential to limiting their impact. It is crucial to assess the accuracy of the current state-of-the-art models for classifying human AI-generated images. In this paper, we look at the Xception model [4]. It has displayed good performance when it comes to accurately classifying images and initially, it has been trained on ImageNet [5]. We analyze how effectively it detects real and manipulated or fabricated images. We look into the performance of the model on different datasets and try to determine if and where it falls short.

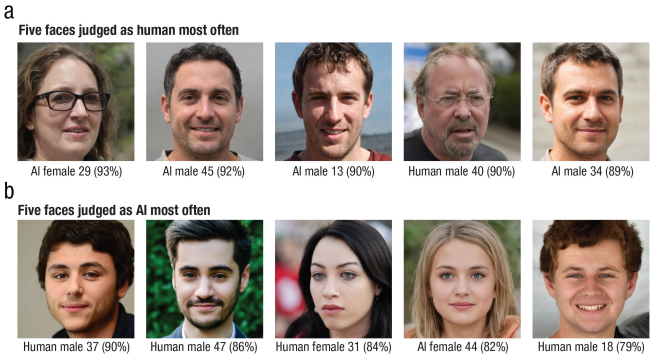

Figure 1: The results of the top five faces most judged as human and AI Courtesy of [1]

The model is only a part of the pipeline for detecting fake images. Examining the shortcomings of the whole process is equally important. We explain why facial detection is not enough, as battling the problem of false images isn't only concerned with facial images, but with people being impersonated in various situations. Scrutinizing the techniques used to detect people and identifying their weaknesses will help pave the way for improving the safeguards for false information.

Examining the various datasets available to train classification models is key as finding places to improve them is instrumental to advancing the field. We look at three datasets that serve distinct functionalities. We believe they can help in advancing the solution to detecting fake images. The first one aims to provide a starting point for detecting hand-manipulated images. These are images that have been altered with Adobe Photoshop[3] or similar programs. The second contains images where the GAN fingerprints [6] have been removed, without changing the RGB representation. This serves to further improve the classifiers, by making it harder for them. The last dataset is thorough, combining images from various facial manipulation techniques, as well as facial synthesis. It collects and generates images from various other works to create a comprehensive dataset. All three datasets assist in improving the current state-of-the-art (SOTA) models.

We want to answer if the Xception model can be relied on to accurately differentiate fake and real faces. In this paper, our contributions are as follows:

- We evaluate existing facial databases used for training detection algorithms, aiming to assist future researchers in making appropriate choices for their situations. Furthermore, this evaluation will aid in identifying limitations and areas for improvement in future work on databases.
- We assess Xception's performance and capabilities

[1]https://github.com/lllyasviel/Fooocus - accessed 27.11.2023

[2]https://www.nvidia.com/nl-nl/geforce/graphics-cards/30-series/rtx-3060-3060ti/

[3]https://www.adobe.com/products/photoshop.html

against multiple datasets

- We discuss the current detection pipelines used in practice and highlight their shortcomings.
- We summarize the set of challenges associated with this issue and provide a set of questions for future researchers.

The Xception model demonstrated great performance across various datasets in the experiments that were conducted. During training with the DFFD dataset, it achieved a validation accuracy of 97.5% and further demonstrated a test accuracy of 97.11%. When evaluated on a subset of 20,974 images from the iFFDB, the model achieved an accuracy of 73.15%, highlighting its potential in real-world scenarios despite some discrepancies due to different training datasets for StyleGAN. Additional tests on subsets from the CASIA and CelebA datasets showed high accuracies of 96.87% and 98.07%, respectively. These results confirm the model's capability to generalize well to different datasets and effectively distinguish between real and fake facial images .

The paper is structured as follows. In Section 2 we talk about the background and the related work. Section 3 we summarise the challenges that current researchers are gacing. We analyse the databases and various detection techniques in Section 4. Section 5 shows the experiments we have done along with the setup and results. We explain why our paper follows the responsible research guideline in Section 6. Finally in Section 7 we conclude our paper and give suggestions on what future researchers can look into.

2 Background & Related Work

In this section we will talk about the background needed to understand the paper and the related work. In Section 2.1 we explain what GANs are and how they work. and Section 2.2 we look at the background knowledge that would be needed to understand the paper fully. In Section 2.3 and Section 2.4 we look at some of the related work that we will examine in the paper

2.1 Generative Adversarial Networks

First introduced by Goodfellow et. al. [7], generative adversarial networks are at the forefront of synthetic image generation nowadays, as they can reliably create highly realistic images. They consist of two neural networks - generator and discriminator, that are pitted against each other to train. Since their inception many advances have been made in the field – Karras et. al. [8]. They progressively grow both the generator and discriminator, meaning that during the training process, they add new layers capturing finer details. Such improvements make it increasingly difficult to distinguish between authentic and fabricated images.

2.2 Image Manipulation Techniques

Properly evaluating the models is crucial to determining the difficulty of detecting facial manipulations. To do that we need balanced datasets that will capture different aspects of face forgery. There are various categories of facial manipulation – face morphing, face swap, 3D modeling, make-up, and face synthesis.

Deep learning techniques are often used to manipulate images [9]. Techniques such as DeepFake[7] have become increasingly popular, in particular with the emergence of FaceAPP[4]. Moreover, GANs are used for full facial synthesis that is indistinguishable from real faces [1]. It is also used for manipulating existing features, such as changing the age, hair, or eye color of people. Face2Face [2] is a real-time reenactment technique used for swapping expressions of different people in videos. It transfers the head position, expression, and blinking from one video to a target video.

2.3 Manipulated Image Databases

iFakeFaceDB [10] is a database containing real and synthetic images, that are pre-processed to remove any GAN fingerprints [6], while not changing the visual appearance.

Hand-crafted fake facial dataset (HFM) [11] is a dataset containing real faces and a set of manually manipulated faces using Adobe Photoshop. The manipulations are with different complexity levels, which improves its variety.

Diverse Fake Face Dataset (DFFD) [12] is the biggest data set we reviewed in this paper, containing 2.6 million facial images. It encompasses all the different techniques mentioned in Section 2.2, making it highly varied and suitable for training models.

2.4 Classifiers and detection methods

Xception [4] is a widely used image classification model. It was developed in 2017 by a Google employee François Chollet. It uses depthwise separable convolutions that factorize a standard convolution into two separate operations: a depthwise convolution and a pointwise convolution. This reduces the computational costs while keeping the expressive power of the model. You can see a Xceptions architecture in Figure 5.

Shallow-FakeFaceNet [11] is a shallow CNN developed such that it can focus on manipulated facial landmarks to detect fake images. The detection pipeline only relies on detecting fake facial images based on RGB information. Being a shallow neural network enables training with less data compared to the state-of-the-art models.

3 Challenges

In this section, we define the challenges associated with the detection of manipulated human images. Different models, such as StyleGAN[5] and Stable Diffusion[6] have different generation approaches making it difficult to detect images from the different models. The varied data used to train different GANs would affect the detectability as some datasets can produce more convincing results than others. Making the classifiers pose invariant, meaning the way they are facing is irrelevant to the result of the classifier. This is more difficult as with the increase of the subclasses more data is needed, making the training process more cumbersome. Moving past recognizing only faces, but rather full body images is another challenge, concerning the lack of data. While real body image datasets

[4]https://www.faceapp.com/

[5]https://github.com/NVlabs/stylegan

[6]https://stability.ai/

exist and are widely used in the automotive industry, synthetic ones are lacking. The creation of new datasets and the tuning of models would prove to be time and power-consuming.

4 Analysis of the databases and detection techniques

In this section, we analyze various databases, models, and detection techniques. We look at their advantages and drawbacks and give out points for improvement. Section 4.1 and Section 4.2 analyze the databases we look at in this paper. In Section 4.3 we look at the Xception [4] model. Finally, the detection pipeline is examined in Section 4.4.

4.1 Synthetic Facial Databases

In order to create a comprehensive database one needs to follow a good framework. This would help in avoiding bias and improving the overall usability of the dataset. Hutchinson et al. [13] mention that a good dataset should be able to answer the 6 W's - Who, What, When, Where, Why, and How. It is crucial to know how the data is addressing the current problem, and if there is a better way to do so. They also emphasize the importance of creating a Dataset Design Document, that lays out how the requirements will be achieved and justifies the decisions that are made. According to Hutchinson et al. [13] when choosing a dataset for your research you need to pay attention to the following characteristics we have summarised:

- Accuracy - How close do the values in the dataset represent the problem?
- Reliability - Do the values contradict each other and are they outdated?
- Consistency - Are the values normalized or resized when it comes to images?
- Size - Is the size of the dataset appropriate for the task at hand?
- Diversity - Does the dataset encompass all the necessary subclasses in order to avoid bias?
- Completeness - Is all the information important for the task and does it fully represent the various subclasses?

Handcrafted Facial Manipulation (HFM)

Handcrafted Facial Manipulation (HFM) [11] is a dataset that contains manually curated faces. It is created by artists with Adobe Photoshop and it consists of 1527 forged images and 621 original images. They are edited in different levels of complexity as can be seen in Figure 2. They have applied various different manipulations to the face such as:

- Attribute manipulation - This is where one or more of the attributes of a person, such as hair, glasses, etc. are changed.
- Face Swap - This manipulation is concerned with swapping the faces of the person with someone else's.
- Attribute swap - This manipulation swaps the attributes of a person with someone else's - changing the mouth or the eye of a person, with someone's.

Lee et al. [11] have sampled various facial images of men and women of different ages and with different facial attributes to improve the diversity of the dataset. Furthermore, they also include complex features such as makeup and beards, as well as images where multiple faces are visible and modified.

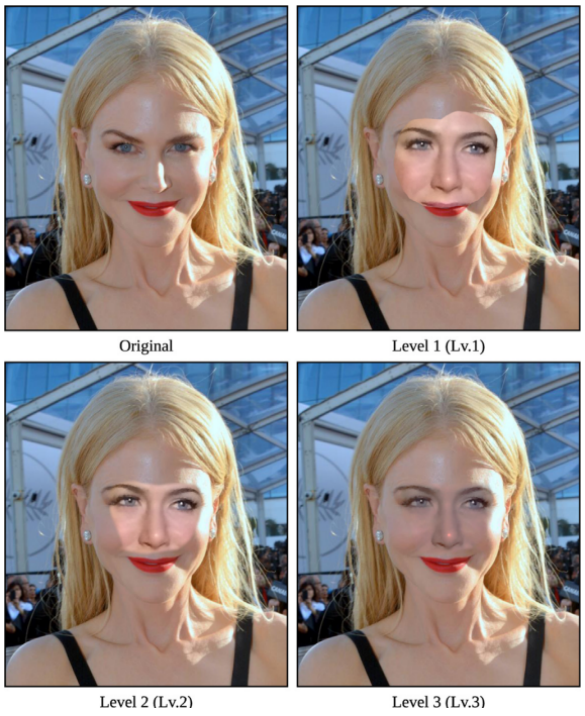

Figure 2: Lv.1 is cropped and pasted, Lv.2 is cropped, pasted, and smoothed edges, and Lv.3 is cropped, pasted, smoothed edges, and adjusted color and light levels. Courtesy of [11]

The issue with the dataset is its size. Depending on the use case it might be too small. For training an image classification model the amount of data is insufficient. This is because such models are typically convolutional neural networks that would need a few thousand images to train the model. It is difficult to give an exact size that is adequate, as it also depends on the extent to which the data accurately represents what might be expected in the real world, rather than its size [14]. In our case 2148 images would be insufficient for training [11]. It is inherently difficult to grow such a dataset as the creation of images is a manual process of artists editing images. This forces scholars to use image augmentation techniques to increase the dataset's size to something usable. The creators of the dataset apply different actions such as image shifting, shearing, zooming, and flipping. However, those techniques have their drawbacks as the new big dataset can be only as good as the initial set of images is, as it cannot create new data, but rather combines or imitates the data [15]. If a class is missing from the initial dataset it cannot be created through data augmentation. Augmenting the fake images can create further complications, as the neural network might learn to recognize augmented data rather than fake or real. This could be mitigated by also augmenting the real set of images, but further research has to be done.

Evaluating it against the characteristics we listed initially,

we can see that it follows them closely. Their accuracy is good, as they represent the expected real-world data very well, through the different levels of editing complexity of the images. The dataset is reliable, as it doesn't contain any contradicting values. Furthermore, the Adobe Photoshop[3] technology hasn't changed much, meaning the data is up-to-date. When it comes to consistency, the dataset is lacking, as the images differ in size. This means that they have to be preprocessed by the user for feeding them to a CNN. The diversity of the dataset is good it includes different races and various facial features, but when it comes to completeness it is lacking, as the limited size doesn't allow for a good representation of the different types of people and features.

iFakeFaceDB (iFFDB)

iFakeFaceDB (iFFDB) [10] is a database compiled from existing databases of synthetic facial images. Its contribution is the GANprintR autoencoder, which removes the GAN fingerprints while not changing the visual representation of the final image. Having images from different databases and differently trained models would make the classification model more robust as it might be able to learn the common patterns of the model. Furthermore, its size is appropriate for adequately training a neural network. Removing the GAN fingerprints makes images harder to recognize by current models, as it removes one dimension that classification models could have exploited. It further pushes models to focus more on the RGB representation of the image rather than its underlying information such as meta data. The drawbacks of the database are inherited from the composite datasets.

When comparing it to the characteristics we listed formerly, we can conclude that this dataset adheres to them. The dataset is accurate and reliable, as the data is sourced from sets with images generated from SOTA GAN models. All images are consistent and resized to 224x224. This allows us to feed a neural network without any unnecessary preprocessing. Moreover, the diversity and completeness of the dataset are good as it consists of people from different races and facial features. Furthermore, it assists in accurately capturing the intricacies of the different facial features and races. The problem is that all the images are front-facing, while in the wild images would differ. That said the dataset serves the purpose of estimating the performance of classifiers. Lastly, the size of the dataset is appropriate for our use case as it allows us to train a neural network.

Diverse Fake Face Dataset (DFFD) [12]

This is the biggest dataset we look at – 2.6 million facial images. It is also very varied as it includes facial manipulations such as identity and expression swap, attribute manipulation, and entire facial synthesis. Examples can be seen in Figure 3. Among all the data 47.7% are from male subjects and 52.3% are female with the majority of people being in the range of 21-50 years of age. For the real facial images, they have used the CelebA dataset [16] and the FFHQ dataset [17]. When it comes to identity swap expression swap and attribute manipulation, they have used general methods such as FaceApp[4], Face2Face [2], Deepfake[7], and others.

[7]https://github.com/deepfakes/faceswap - accessed 11.09.2019

Figure 3: Different types of facial manipulation - Courtesy of [12]

Evaluating this dataset to the characteristics we listed initially, we see that it has some issues when it comes to reliability and accuracy. Using different publicly available models and techniques for manipulating images would make the dataset more accurate as it is a better representation of what a malicious user might use. However, technology progresses and the techniques used to create those images are advancing more more. For instance, StyleGAN [18] has had updates over the years improving on the technology. This doesn't mean that the dataset is obsolete, yet, but with time it will become a more and more inaccurate representation of real-world data. The dataset is consistent as all the images they provide are scaled to a ratio of 299x299, streamlining the training process. The diversity and completeness of the dataset are further improved by the size making sure different sub-classes are appropriately represented.

4.2 Real Facial Databases

For the real facial databases, we have choses CelebA and CASIA-WebFace. Firstly, both datasets provide high-quality images that accurately represent the wide range of facial features and variations found in real-world scenarios, ensuring that the values in the datasets are close to the actual problem of distinguishing real faces from synthetic ones. The reliability of CelebA and CASIA-WebFace is also a significant advantage. These datasets are carefully curated and widely used in the research community, ensuring that the data is up-to-date and free from contradictions. Consistency in the datasets is another strong point. The images are normalized and resized, which helps maintain uniformity across the training data. Furthermore, it requires less pre-processing before training. Moreover, the size of these datasets is appropriate for the task at hand. CelebA contains over 200,000 images, while CASIA-WebFace includes over 490,000 images, providing a substantial amount of data for training convolutional neural networks. This large volume of data is crucial for the deep learning models to learn complex patterns and generalize well to new data. Diversity is another critical aspect where these datasets excel. They encompass a wide range of facial variations, including different expressions, poses, and lighting conditions, as well as a broad spectrum of demographic attributes. This diversity helps in avoiding biases and ensures that the models trained on these datasets are more inclusive and generalizable. Finally, the completeness of CelebA and CASIA-WebFace is notable. They provide comprehensive information that fully represents various subclasses within the domain of real facial images.

To summarize this section, creating a comprehensive database requires adherence to key qualities such as accuracy, reliability, consistency, size, diversity, and completeness. The

Handcrafted Facial Manipulation (HFM) dataset, despite its small size and need for augmentation, provides diverse and accurate representations of hand-made facial manipulations, but it is not suitable for training neural networks without the assistance of data augmentation. The iFakeFaceDB (iFFDB) offers a big dataset of synthetic facial images, that have been further improved by removing the GAN fingerprints. This together with its size, makes it highly suitable for training deep-learning neural networks. The Diverse Fake Face Dataset (DFFD) stands out with its extensive size and varied manipulations. We have chosen DFFD and iFFDB for our experiments, as their size makes them great for training CNNs.

4.3 Models

Xception is one of the most widely used image classification models. It stands for "Extreme Inception", a successor of the Inception model [19]. It substitutes the standard Inception modules with depthwise separable convolutions. Inception modules are similar to convolutions, as they can learn complex representations, a simplified one can be seen in Figure 4. Deptwise separable convolutions factorize a standard convolution into two separate operations: a depthwise convolution and a pointwise convolution. The former performs a single convolution on each input channel and the latter uses a 1x1 convolution to combine those outputs. This reduces the computational costs in terms of time and memory, as there are significantly fewer parameters while keeping the expressive power of the model. You can see the Xception architecture in Figure 5.

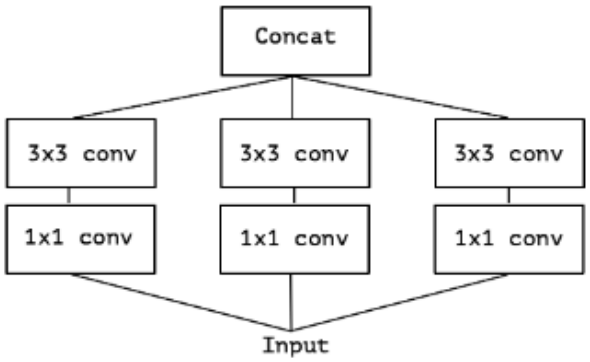

Figure 4: Simplified structure of the Inception module courtesy of [4].

Xception has shown promising results when it comes to detecting facial manipulations compared to other models, as can be seen in Figure 7 and the papers [12] and [20]. It is considered a state-of-the-art model for image classification as it is the most common benchmark for performance in the literature we have looked at in the paper.

We can observe in Figure 7 that with the decrease of the quality the performance of the models decreases. The efficacy drops when it comes to low-quality and hand-crafted images [9]. Shallow Convolutional Neural Networks(CNNs) also experience a drop in accuracy according to them.

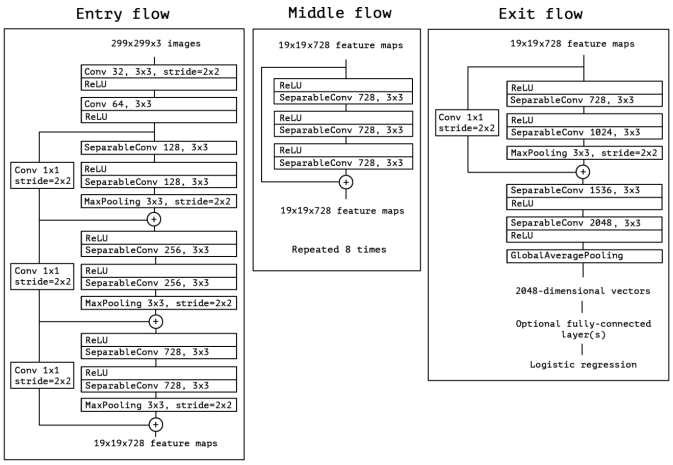

Figure 5: The Xception architecture: the data first goes through the entry flow, then through the middle flow which is repeated eight times, and finally through the exit flow. Note that all Convolution and SeparableConvolution layers are followed by batch normalization (not included in the diagram). All SeparableConvolution layers use a depth multiplier of 1 (no depth expansion). Courtesy of [4].

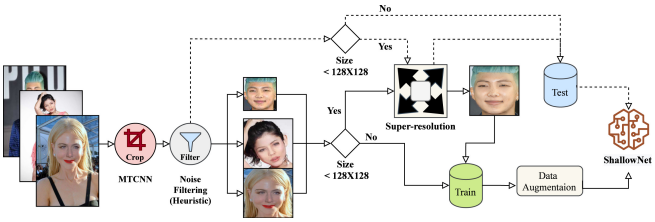

Figure 6: Shallow-FakeFaceNet end-to-end pipeline to detect facial manipulated images. Courtesy of [11].

4.4 Detection pipelines

While the classification model is important, we also have to look at the whole process of detecting a false image. The goal is to determine if the detection techniques are performing well in real-world scenarios. In Figure 6 we can see the detection pipeline of the Shallow-FakeFaceNet. It precisely illustrates that pipeline and we can see where it can be improved. They are using a Multi-Task Cascaded Convolutional Neural Network (MTCNN) to extract the faces from the images. The benefit of doing so is it allows us to train a classifier on faces, which we can then use for all sorts of images, as long as we can see the face of the person. The drawback is that an image might have a genuine face but a manipulated body. For instance, an image of a public figure can be manipulated to show them in an inappropriate pose, to hurt their credibility for example. The use of MTCNN for extracting faces from images for training and detecting can also be seen in [20], [21]. While the detection of facial manipulations is crucial to stop malicious users from exploiting people, we want to urge researchers to go into the detection of the whole body. Currently, there exist databases of real people in various poses, as they are used in the automotive industry for detecting pedestrians, such as the Human Dataset V2 [22]. The problem is with the lack of synthetic datasets. Their creation would

Compression	Raw	HQ	LQ
[14] XceptionNet Full Image	82.01	74.78	70.52
[27] Steg. Features + SVM	97.63	70.97	55.98
[17] Cozzolino *et al.*	98.57	78.45	58.69
[10] Bayar and Stamm	98.74	82.97	66.84
[51] Rahmouni *et al.*	97.03	79.08	61.18
[5] MesoNet	95.23	83.10	70.47
[14] XceptionNet	**99.26**	**95.73**	**81.00**

Figure 7: Accuracy of different models. HQ (constant rate quantization parameter equal to 23) which is visually nearly lossless. Low-quality videos (LQ) use a quantization parameter of 40. Images are conservatively cropped around the center of the tracked face. XceptionNet Full image is a baseline benchmark. Courtesy of [20]

require more time and processing power than we have now, which is why we are leaving it to future researchers.

5 Experiments

In this section, we will look at the experiments we have performed. We aim to measure the Xception model's accuracy [4] and see how it performs on unseen datasets. Based on the studies we have looked at [10], [11], [12] we are expecting results up to the 90th percentile. As some of our tests are performed with lower quality images we are expecting worse accuracy on them. This section is structured as follows. Section 5.1 shows the model and process for the experiments as well as the hyperparameters and the databases we are using. In Section 5.2 we explain the setup for each experiment, namely how the amount of data used, how it has been pre-processed, and the train/validation/test split. Finally, our results and conclusions can be seen in Section 5.3

5.1 Experiment Tools

Our experiments have been performed on the Xception model [4], pre-trained on the ImageNet dataset [4]. We have split the training process into two stages - pre-training where we freeze the body layers, and regular, where all the layers can be adjusted, similar to the training process seen in [20]. This assists in a faster convergence and reduces overfitting.

Parameter	Value
Number of epochs (pre-training)	5
Number of epochs (fine-tuning)	15
Batch size (pre-training)	32
Batch size (fine-tuning)	16
Learning rate (pre-training)	1e-3
Learning rate (fine-tuning)	1e-4

Table 1: Hyperparameters for training stages

For the fake faces we have used iFakeFaceDB[10] and DFFD[12], whereas for real facial images, we have used CelebA[16] and CASIAWebFace[23]. We decided on those two real image databases as they are widely used in the field. Depending on the database and the experiment we have pre-processed the images differently. For the training, we used a publicly available GitHub repository[8] with the Xception Model from Keras[9]. The hyperparameters we have used are taken from the aforementioned repository and can be seen in Table 1. We have modified the implementation to suit our needs and our code can be found in our GitLab repository[10]. It also contains the code we have used for extracting, cropping, aligning, and compressing the images to normalize them for the neural network.

5.2 Experimental Setup

DFFD

For this experiment, we downloaded the DFFD dataset[11] and used a subset of the data to train the Xception model [4]. We used the provided subset of Flickr-Faces-HQ(FFHQ) [18] for the real images and for the fake images we used the set they provided created by StyleGan [18] trained on FFHQ, refer to Figure 10. For this experiment, there was no pre-processing required, as the images were aligned and had the same resolution. The number of epochs for fine-tuning is 8. The total amount of images for training and validation is 21,998 out of which 10,999 are fake and 10,999 are real. We used an 85%/15% split for the training and validation. Furthermore, our test set contained 8997 fake images from generated by StyleGAN and 9000 real images from FFHQ. The experiment ran for 5 hours on the DelftBlue supercomputer, using 8 CPUs with 16GB of memory per CPU.

CASIAWebFace & iFakeFaceDB

For this experiment, we had to pre-process the images to avoid any bias. The images were differing, as iFFDB contained only cropped-out forward-looking faces, while CASIA had more variety. When extracting we cropped and aligned the faces such that they are looking forward. We followed the same methodology explained in [10]. We are extracting 68 face landmarks with the technique described in [24]. Kazemi and Sullivan's implementation is used in dlib's[12] pose predictor, and we downloaded the pre-trained weights for it. Since the images were in different resolutions we decided to normalize all of them to 112x112. For that, we used openCV[13]. The images from the CASIA dataset were of lower quality and instead of upscaling them, we decided to compress the images from iFFDB. We did this to avoid any possible artifacts being introduced from upscaling. The total amount of images for training and validation is 36,680, where the train/validation split was 80%/20%. The fake faces are 20,000 and the real faces are 16,680. For testing, we chose 3000 images of both classes from their respective databases. The experiment ran for 4 hours on the DelftBlue supercomputer, using 8 CPUs with 16GB of memory per CPU.

[8] https://github.com/otenim/Xception-with-Your-Own-Dataset

[9] https://keras.io/

[10] https://gitlab.ewi.tudelft.nl/cse3000/2023-2024-q4/Lukina/fdobrev-The-Many-Faces-of-AI-Art.git

[11] https://cvlab.cse.msu.edu/dffd-dataset.html

[12] http://dlib.net/

[13] https://docs.opencv.org/

Figure 8: Examples of unsuitable images that passed through our pre-processing stage.

It is important to mention since the images were not manually curated, some unsuitable and wrongfully detected images, see Figure 8 have gone through our filters, but due to their limited number, we believe it won't affect our results.

CelebA & iFakeFaceDB

This experiment required pre-processing due to the difference in the size and types of the images. We had to crop and align the images from CelebA to make them similar to iFFDB. Further more we normalized them to 112x112 for the same reasons as mentioned in Section 5.2. Our training and validation set consisted of 39,995 images - 19,995 of which were real from the CelebA database and 20,000 fake were taken from iFFDB. The train/validation split was 80%/20%. Since we had big databases on our hands we decided to create a huge test set. We chose 20,006 different real images at random from CelebA and 20,974 different fake images from iFFDB. This would allow us to get a better idea of the performance of the model. The experiment ran for 4 hours on the Delft-Blue supercomputer, using 8 CPUs with 16GB of memory per CPU

5.3 Experimental Results

DFFD Results

The Xception model achieved 97.5% validation accuracy during training. Furthermore, it achieved 97.11% accuracy on our test set. In Figure 9 we can see how the accuracy changes with each epoch. In the first five epochs during the pre-training, we observe that the accuracy is just below 60%. After the full training of the model commences we can see that it quickly converges to 97% accuracy. This could be because we are using the model with initialized weights on ImageNet, allowing us to use the patterns it has found before. Further testing on the hyperparameters needs to be done to identify the reason for the fast convergence. Moreover, to test the model's applicability in the real world, we tested it on the subset of 20,974 images from iFFDB we used for our other experiments. It achieved an accuracy of 73.15%, which shows that the model can perform reasonably on different datasets. More of the results can be seen in Figure 11. It is important

Figure 9: Validation Loss and Validation Accuracy for the Xception model trained on images from DFFD

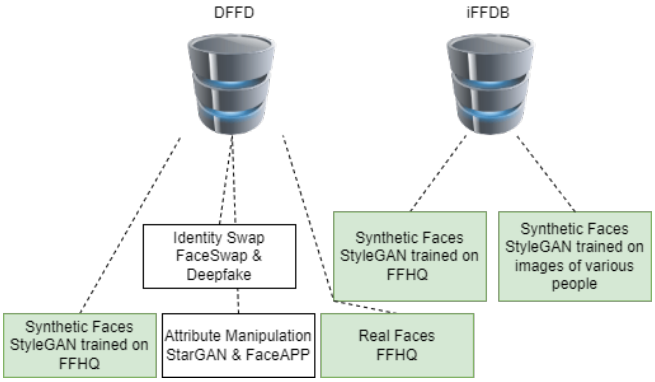

Figure 10: The contents of the two fake databases we have considered, and the parts we have used are colored in green

to mention that the subset we used for testing has also been generated with StyleGAN[18], similar to DFFD, see in Figure 10. The difference to DFFD is that part of the images in iFFDB, are generated by StyleGAN [18] trained on different data. The lower accuracy plays into one of the mentioned challenges, namely that the same generative model trained on different datasets can produce different results that are hard to generalize. This discrepancy in accuracies can also be attributed to the removal of the GAN fingerprints from the images in the iFFDB. We also tested the model on two subsets of 3,000 images from CASIA and CelebA datasets, achieving a 96.87% accuracy and 98.07% accuracy accordingly. This serves as proof of the capabilities of the model to recognize real and fake facial images from outside datasets.

5.4 CASIAWebFace & iFakeFaceDB

After training the model, we achieved a validation accuracy of 0.9999. This was initially a surprising result that might lead us to conclude that the model is overfitting. Contrary to that when we ran the test suite of 6,000 images split 50/50 between the classes it incorrectly identified only 2 of them, meaning an accuracy of 0.9996. Looking at the results of Neves et. al. [10] we can see that they also achieve an accuracy of 99.9+% on the TPDNE and 100F fake datasets with

Training & Validation Data		Testing Data		Accuracy	
Real	Fake	Real	Fake		
DFFD(FFHQ) - 10,999	DFFD - 10,999	DFFD - 9,000	DFFD - 8,997	97.11%	
DFFD(FFHQ) - 10,999	DFFD - 10,999	CelebA - 3,000	iFFDB - 20,974	98.07%	73.15%
DFFD(FFHQ) - 10,999	DFFD - 10,999	CASIA - 3,000		96.87%	
CASIA - 16,680	iFFDB - 20,000	CASIA - 3,000	iFFDB - 3,000	99.96%	
CASIA - 16,680	iFFDB - 20,000	DFFD - 9,000	DFFD - 8,997	50.01%	
CelebA - 19,995	iFFDB - 20,000	CelebA - 20,006	iFFDB - 20,974	100.00%	
CelebA - 19,995	iFFDB - 20,000	DFFD - 9,000	DFFD - 8,997	52.30%	

Figure 11: Aggregated results of all experiments. We can see the different datasets used for training and testing, together with the amount of images used next to them. The overall accuracy is displayed in the rightmost column. For the second row we have decided to split the accuracy in two to show the accuracy on real(green) and on synthetic(red) facial images

CASIA used for the real images. However, those results are achieved before applying the fingerprint removal, which in our case is applied. According to them, our results should be in the range of 95%. This result is unexpected and we could not pinpoint the exact reason. We tested the model on unseen data from DFFD but the achieved accuracy was no better than random guessing, as seen in Figure 11. One possibility is that the images weren't pre-processed accordingly, which could have been easy for the neural network to capture. Neves et al. [10] did not provide the code for aligning and extracting the frontal faces, so we followed their instructions as shown in Section 5.2. Another reason could have been the quality of the facial images. CASIA's images were noticeably worse than in iFFDB even though we were working with 112x112 images. This was the reason we decided to switch the datasets and use CelebA.

5.5 CelebA & iFakeFaceDB

The results from this experiment were the same as with CASIA. The validation accuracy was again 0.9999. Upon running it on our test set we get 100% accuracy, as all 40,980 images were correctly identified. The fact that it correctly labeled all the images means that there should exist an intrinsic difference between them. Here we also tested the model on unseen data from DFFD achieving an accuracy of 52.30%, slightly better than random guesses, as seen in Figure 11. Since the image cropping and alignment are not available in their GitHub repository[14], the possibility of making mistakes in this part of the process is real. Furthermore, there exists the possibility that the iFakeFaceDB has an intrinsic fingerprint that is easy to detect by neural networks, but this should be examined by future researchers.

Despite its high performance, the Xception model has several downsides. First, it can be computationally intensive, requiring significant resources for both training and inference. This can make it less suitable for deployment in resource-constrained environments. Second, while it performs well on known datasets, its accuracy can drop when tested on datasets with different characteristics, such as images generated by different versions of generative models. This indicates a potential issue with generalization.

[14]https://github.com/joaocneves/gan_fingerprint_removal

6 Responsible Research

In this section we explain how our research adheres to the guidelines of responsible research. Section 6.1 Explains the ethical concerns we pay attention to during our research. In Section 6.2 we show that our research is reproducible. Finally, we explain how we avoid plagiarism and bias in Section 6.3.

This research work is compliant with Chapters 2 and 3 from the Netherlands Code of Conduct [20], as conducted by Filip Dobrev, a student at Technische Universiteit (TU) Delft, under the supervision of Anna Lukina

6.1 Ethical concerns

The ethical considerations in AI-generated face detection research are important, given the potential impacts on privacy, consent, and societal trust. One primary concern is the privacy of individuals whose images are used in datasets. Ensuring that all images, whether real or synthetic, are sourced with proper consent is crucial. We respect privacy rights and we use only publicly available datasets, that are standard for the industry. Additionally, addressing potential biases in datasets and detection models is essential to develop fair AI models that do not discriminate or misidentify specific demographic groups. It is important to mention that people with skin conditions skin conditions such as vitiligo or rosacea are underrepresented leaving the possibility of the models favoring against them [11]. Limiting the biases in the training process assists in mitigating the risks of misuse. In Section 4 we explain in detail what the datasets consist of and any limitations they have. We do not work with or process any personal data that is not available to the public.

Finally, the societal impact of AI-generated face detection research cannot be overlooked. While these technologies can enhance security, they also raise concerns about privacy and consent. People might not want their images to pass through an AI detection system. The benefits of detection technologies should realized without compromising individual freedoms and societal norms.

6.2 Reproducibility

Maintaining transparency in the research process is equally important. We are documenting and openly sharing methodologies, datasets, and model architectures to enable peer review and replication of results, fostering trust and ensuring accountability. We also communicate the limitations of detection technologies clearly. Furthermore, we provide an explanation of the experimentation setup and the experiments themselves in Section 5

6.3 Plagiarism

We are committed to keeping scientific integrity, by avoiding plagiarism and conflicts of interest. To achieve that we document the sources we have used, and we actively reference them, to keep transparency. Additionally, we have no affiliations with the authors of the papers cited and do not receive any financial compensation, thus preventing conflicts of interest.

7 Conclusions and Future Work

This research demonstrates the high accuracy of the Xception model in detecting AI-generated faces across various datasets, including DFFD, CASIAWebFace, CelebA, and iFakeFaceDB. The model achieved near-perfect validation accuracy, suggesting a significant intrinsic difference between real and synthetic images. Despite concerns about overfitting, the consistent results across multiple tests indicate the robustness of the Xception model in this domain. However, the potential existence of dataset-specific fingerprints that are easily detectable by neural networks calls for further investigation. This study highlights the necessity for diversified and comprehensive datasets to avoid biases and ensure the generalizability of detection models.

Future research should focus on several key areas to advance the field of AI-generated image detection. First, exploring different models and optimizing hyperparameters could enhance detection accuracy and robustness. Additionally, addressing the potential biases in datasets by including underrepresented groups, such as individuals with skin conditions and facial deformities, is crucial for developing fair and unbiased classifiers. Investigating the existence and impact of dataset-specific fingerprints could provide insights into improving model generalization. Furthermore focusing on detecting humans more generally, rather than faces only should be the next step. Finally, expanding the scope of research to include real-time detection in social media and broader applications of these models will be essential for combating the evolving challenges posed by synthetic media.

References

[1] E. J. Miller, B. A. Steward, Z. Witkower, C. A. M. Sutherland, E. G. Krumhuber, and A. Dawel, "Ai hyperrealism: Why ai faces are perceived as more real than human ones," *Psychological Science*, vol. 34, no. 12, pp. 1390–1403, 2023, pMID: 37955384. [Online]. Available: https://doi.org/10.1177/09567976231207095

[2] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," 2020.

[3] B. Contreras, "Tougher ai policies could protect taylor swift—and everyone else—from deepfakes," *Scientific American*, 2023. [Online]. Available: https://www.scientificamerican.com/article/tougher-ai-policies-could-protect-taylor-swift-and-everyone-else-from-deepfakes/

[4] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.

[5] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Neural Information Processing Systems*, vol. 25, 01 2012.

[6] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do gans leave artificial fingerprints?" 2018.

[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

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

[9] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection," 2017.

[10] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proenca, and J. Fierrez, "Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, p. 1038–1048, Aug. 2020. [Online]. Available: http://dx.doi.org/10.1109/JSTSP.2020.3007250

[11] S. Lee, S. Tariq, Y. Shin, and S. S. Woo, "Detecting handcrafted facial image manipulations and gan-generated facial images using shallow-fakefacenet," *Applied Soft Computing*, vol. 105, p. 107256, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494621001794

[12] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain, "On the detection of digital face manipulation," 2020.

[13] B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell, "Towards accountability for machine learning datasets: Practices from software engineering and infrastructure," 2021.

[14] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A. B. Dris, N. Alzakari, A. Abou Elwafa, and H. Kurdi, "Impact of dataset size on classification performance: An empirical evaluation in the medical domain," *Applied Sciences*, vol. 11, no. 2, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/2/796

[15] B. Hüttenrauch, *Limitations of data augmentation and outlook*. Wiesbaden: Springer Fachmedien Wiesbaden, 2016, pp. 279–290. [Online]. Available: https://doi.org/10.1007/978-3-658-14577-4_8

[16] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[17] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2019.

[18] ——, "A style-based generator architecture for generative adversarial networks," 2019.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014.

[20] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," 2019.

[21] J. El Abdelkhalki, M. Ben Ahmed, and A. A. Boudhir, "Deepfake detection based on the xception model," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, pp. 400–406, 2021.

[22] H. v2, "Human dataset v2 dataset," https://universe.roboflow.com/human-v2/human-dataset-v2, apr 2022, visited on 2024-06-08. [Online]. Available: https://universe.roboflow.com/human-v2/human-dataset-v2

[23] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014.

[24] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.