ARE YOU REAL? DISTINGUISHING GENUINE FACES FROM AI

Maja Joanna Gojska (s243077), Claudio Eugenio Castorina (s243076),
Inaki Zabaleta (s240056) and Joel Farre (s240076)

Technical University of Denmark
DTU Compute

ABSTRACT

The rapid progress of generative AI facilitated the creation of highly realistic human images, making it difficult to differentiate genuine photographs from AI-generated fakes. This study evaluates the effectiveness of machine learning and deep learning models in classifying real versus synthetic human images. We assembled a dataset of approximately 40,000 images, equally split between real photos from Flickr and synthetic images generated using StyleGAN, StyleGAN2, and Stable Diffusion 1.2. We developed an advanced model that combines features from a pretrained EfficientNetV2 convolutional neural network (CNN) with a multilayer perceptron (MLP) that processes color distributions and wavelet energies. This model reached test accuracies of around 96.6% and successfully generalized to out-of-distribution images from DALL-E 3, Stable Diffusion 2.0, and Midjourney. Explainability techniques like SmoothGrad and GradCam showed that the model effectively uses higher-level facial features for classification. These findings demonstrate the model's strong ability to detect AI-generated human images, offering potential solutions to mitigate the misuse of generative technologies.

***Index Terms*—** GANs, Deep Fake, Image Classification, Deep Learning, Explainable AI, Image Authentication, Out-of-Distribution Generalization

1. INTRODUCTION

1.1. Background and Motivation

The rise of generative artificial intelligence has made it increasingly simple to create hyper-realistic images of humans. For instance, a study by *Nightingale et al.* demonstrated that images generated by StyleGAN2 are advanced enough to not be discernible by humans [1]. Similarly, You et al. conducted an evaluation of images generated by Stable Diffusion, arriving at comparable conclusions [2].

The widespread of AI-generated content raises significant concerns, particularly in the context of fake faces. While many generative AI applications are used responsibly, fake face images have been associated with a range of potential misuse scenarios, including identity fraud, misinformation campaigns, and social engineering attack [3]. The simplicity with which such images can be generated and distributed online makes them a critical area of study.

1.2. Research Questions and Objectives

Against this background, the research questions that will guide the investigation are:

1. How consistently can we tell real and fake people apart (in terms of accuracy)?
2. Can the model generalize to images generated from a different distribution?
3. Can we explain (some of) the decisions of the model?

By addressing these questions, this study aims to contribute to the development of robust, generalizable, and interpretable deep learning solutions for detecting fake faces.

2. RELATED WORK

Recently, many methodologies have been developed to distinguish between real and fake images. Traditional approaches often rely on image processing techniques such as wavelet decomposition [4], analyzing color distribution differences [5], and detecting traces of demosaicing filters [6]. However, these methods tend to be highly dataset-specific and may lack robustness across diverse datasets. Although these methods can be effective, they are typically highly specific to particular datasets and may struggle to maintain robustness across diverse image collections.

Deep learning has become the primary technique for fake image detection, particularly through the use of convolutional neural networks (CNNs). CNNs demonstrate high performance in identifying manipulated images by learning intricate patterns and features that traditional methods might overlook [7]. For instance, the ViXNet framework introduced by [8] used a vision transformer combined with the Xception network to identify forged and swapped faces. Similarly, [9] presented an enhanced dense CNN framework trained on various deepfake image datasets to improve generalization capabilities. However, this framework showed limited performance when applied to unseen datasets.

The research field focused on human faces is less extensive compared to general fake image detection. *Patel et al.* proposed a method for detecting fake images generated with an autoencoder-based face swap, which estimates 3D head poses from face images to achieve high performance [10]. Another study evaluated the generalization capabilities of both deep learning and texture-based methods for fake face detection, finding that while CNN-based methods achieve high accuracy on their training datasets, their performance significantly declines with fake face generation techniques not included in the training data [11]. This proves the difficulty in achieving robust performance in OOD detection scenarios.

The explainability of deepfake detection models is an important area that has received limited attention. Understanding the decision-making processes of deep learning models can provide valuable insights and increase trust in their predictions. Techniques such as Local Interpretable Model-Agnostic Explanations (LIME), Layer-Wise Relevance Propagation (LRP), and analysis of intermediate activations have been used to interpret fake image detection models [12]. These approaches help illustrate how models differentiate between real and fake images, although comprehensive research in this area remains sparse.

While significant progress has been made in the detection of fake images and faces using deep learning, challenges remain in areas such as OOD detection, generalizability, explainability, and the development of effective unsupervised methods. This study aims to adress these issues, thus contributing to the broader effort of mitigating the misuse of Generative AI.

3. METHODOLOGY

3.1. Dataset

The study utilized a dataset comprising approximately 40,000 images, evenly divided between real and synthetic samples. The real images were sourced from Flickr, curated by NVIDIA [13]. Synthetic images were generated using StyleGAN [13], StyleGAN2 [14], and Stable Diffusion 1.2 [15].

To minimize potential biases inherent in the dataset, both the original images and their background-removed counterparts were analyzed. Background removal was performed using Google's SelfieSegmentation tool [16]. The dataset was then partitioned into three subsets: approximately 26,000 images were allocated for training, 8,000 for validation, and 9,000 for testing. Each of these set is evenly split between the two classes. The training and validation split were augumented using the transformation described in the Appendix.

To illustrate the distribution of the dataset, we employed a non-linear dimensionality reduction method (UMAP) [17] to project the images into a two-dimensional space. The resulting visualization is shown in Figure 1.

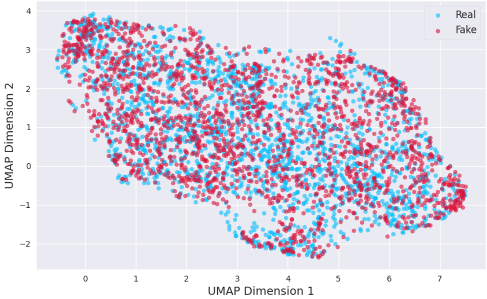

Fig. 1. Two-dimensional UMAP embedding of the original dataset.

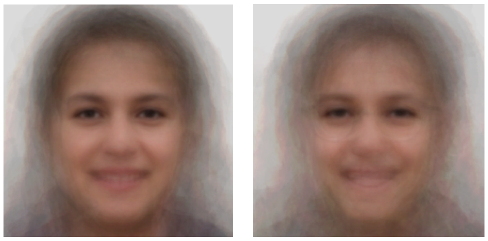

Fig. 2. Superimposion of fake and real faces, respectively.

3.2. Feature Extraction

To assess potential patterns and artifacts that might serve as shortcuts to distinguish AI-generated (fake) images from real ones, we conducted a series of exploratory analyzes focusing on visual characteristics and statistical differences in color and frequency domains.

To investigate whether recurrent structural details could explain classification performance, we visually inspected the spatial alignment of images through superimposition. Fake and real images were overlaid separately to evaluate potential similarities or artifacts that could be used as shortcuts by the classification model. Figure 2 illustrates the superimposed results for both categories. Although general differences in feature distribution were noticeable, no dominant recurring patterns were identified. This warranted further investigation into color and frequency-based characteristics.

Previous work by *Li et al.* [18] has highlighted the predictive value of color distributions in distinguishing AI-generated images. Building on this, we computed color histograms for each channel in the RGB color space. To quantify the statistical significance of the differences between real and fake images, we applied the Kolmogorov-Smirnov (K-S) test to compare distributions.

Inspired by Dieleman [19], we also examined differences in the frequency domain by analyzing wavelet energy at multiple decomposition levels. Wavelet transforms allow us to capture variations in texture and fine-grained patterns that might distinguish generated images.

For each decomposition level, we calculated wavelet energy values and performed independent two-sample t-tests to evaluate statistical differences. The results were highly significant both with and without backgrounds, reinforcing the hypothesis that frequency-domain features can effectively capture discrepancies between real and generated images.

Dataset	Color	Wavelet
Without Background	$< 1 \times 10^{-8}$	≈ 0
With Background	0.029	0.007

Table 1. p-values for color distribution and wavelet energies.

These preliminary results validated the relevance of our feature selection and motivated the need for a more complex architecture capable of leveraging both spatial and structural patterns within the images.

3.3. Baseline

To establish a performance baseline for our supervised learning model, we first evaluated simpler approaches to determine the minimum threshold the CNN should surpass to be considered a success. Specifically, we employed XGBoost, a ML model notorious for high performance with tabular data [20], with the two engineered features, color distribution and wavelet energies.

3.4. Model Architecture

For the supervised learning task, we developed a model comprising three components:

1. An EfficientNetV2-s CNN model pretrained on ImageNet and fine-tuned using the augmented image dataset to extract high-level visual features. The choice of the model is motivated by its state-of-the-art performance-efficiency ratio.
2. A multilayer perceptron trained on the two engineered features, color distribution and wavelet energy coefficients, to capture relevant statistical and frequency-domain information.
3. A fully connected layer combining the learned representations from the previous components to make the final decision.

Figure 3 provides a detailed schematic of each component within the model architecture.

4. EXPERIMENTAL SETTING

4.1. Generalization

To assess the model's ability to generalize beyond the training data's known distributions, we conducted Out-Of-Distribution (OOD) detection experiments. Specifically, we evaluated the model using images generated from three state-of-the-art generative models:

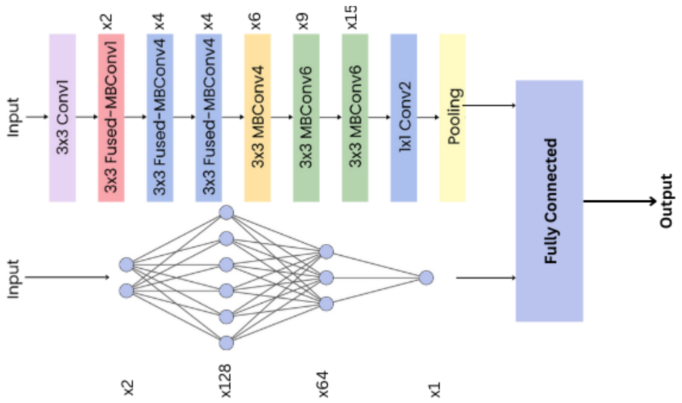

Fig. 3. Architecture of the neural network.

- DALL-E 3, a powerful text-to-image generation model developed by OpenAI, capable of producing highly detailed and diverse images based on textual descriptions.
- Stable Diffusion 2.0, An advanced diffusion model known for generating high-quality images with intricate details and varying styles.
- Midjourney, a proprietary AI image generation tool that specializes in creating artistic and stylistically rich images from textual prompts.

4.2. Explainability

To interpret the outputs of our model, we employed two widely recognized techniques for generating saliency maps:

- **SmoothGrad**: averages the gradients of the image with respect to the output class over multiple noisy copies of the input image.
- **Grad-CAM**: generates localization maps by utilizing the gradients of the target class flowing into the final convolutional layer.

5. RESULTS

The baseline XGBoost model exhibited moderate classification performance across different dataset conditions. Specifically, when trained on the dataset containing background features, the model achieved an accuracy of 69%. However, upon removing the background features, the model's accuracy improved significantly to 81.5%. Table 2 summarizes the accuracy metrics under both conditions.

Dataset	Accuracy
Without Background	0.815
With Background	0.69

Table 2. XGBoost performance on the two datasets.

In contrast to the baseline model, our composite model demonstrated substantial improvements in classification accuracy. The model achieved an accuracy of approximately 96.6% to 96.8% across both datasets with and without background features. Furthermore, when evaluated on an Out-Of-Distribution (OOD) dataset, the composite model maintained a high accuracy of 96.7%. Table 3 presents the detailed performance metrics of the composite model.

Dataset	Test Accuracy	OOD Accuracy
With Background	0.9658	0.9667
Without Background	0.9682	0.9667

Table 3. Composite model performance on the three datasets.

6. DISCUSSION

Our primary goal was to develop and evaluate a method for distinguishing real human faces from those generated by advanced generative models, and to test the generalizability of the approach on images generated by different probability distributions. The results demonstrate that the chosen approach is both effective and generalizable, while also providing insights into the interpretability of model predictions.

The composite model demonstrated a strong capacity to distinguish real faces from fake ones, achieving accuracies of approximately 97% across both datasets (with and without background). The performance significantly surpasses the XGBoost baseline, which showed accuracies of 69% and 81%, for with and without background, respectively. We hypothesize that the lack of significant differences between the two datasets in the performances of the composite model suggest that CNN-extracted features are carrying the most feature importance. This hypothesis is consistent with the observations in earlier studies [7][9], which show comparable performance (In-Distribution) using only CNN models.

The model performed exceptionally well on out-of-distribution (OOD) images generated by unseen architectures, achieving nearly identical accuracy levels (96.67%) as observed in the in-distribution test sets. We propose that the observed performance is mainly driven by aggressive data augmentation and cross-domain feature extraction. Future research should investigate this hypothesis to confirm its validity.

Finally, interpretability of the model was explored using SmoothGrad and Grad-CAM techniques to generate saliency maps of the predictions. The visualizations indicated that the model primarily focused on high-detail facial regions, such as eyes, nose, and mouth (Figure 4.1), which are known to be challenging for generative models to replicate perfectly [21]. Additionally, the model demonstrates considerable general world knowledge. For instance, Figure 4.2 shows that the model recognizes that the left half of the glasses are missing. This behavior aligns with human intuition and prior literature, strongly suggesting that the model relies on task-relevant features rather than superficial patterns, such as background artifacts.

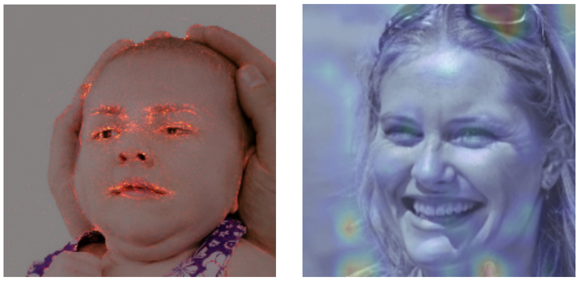

Fig. 4. SmoothGrad saliency map and Grad-CAM visualization of selected fake images.

Despite these promising results, certain limitations remain. The proposed model is inherently temporary; as generative models evolve, this model could potentially be misused as a discriminator in a GAN, which would learn how to bypass it. Additionally, while the images from the OOD testing come from state-of-the-art proprietary models, it is possible that these models share many similarities with the models employed in the training set. Future research should focus on implementing techniques to resist adversarial attacks, such as gradient obfuscation and output perturbation. Additionally, developing unsupervised or self-supervised methods to enhance adaptability, along with novel approaches to improve interpretability—such as activation maximization and counterfactual explanations—would be beneficial.

7. CONCLUSION

The results demonstrate that the proposed hybrid detection approach can reliably distinguish real faces from synthetically generated ones, achieving near-perfect accuracy. In particular, we found that:

1. The model successfully differentiated between real and AI-generated human faces with an accuracy of 97%.

2. The model maintained a high classification accuracy of approximately 97% when applied to out-of-distribution (OOD) images generated by previously unseen models.

3. Although not consistently across all instances, saliency maps generated using SmoothGrad and Grad-CAM techniques effectively highlighted important facial regions (e.g., eyes, nose, mouth). These visualizations provide partial validation of the model's decision-making by focusing on biologically and visually significant features.

These findings indicate that the proposed method effectively detects synthetically generated faces with high accuracy and strong generalizability, offering a robust tool for synthetic image identification and contributing to the ongoing effort to combat the misuse of generative models.

8. REFERENCES

[1] Sophie J. Nightingale and Hany Farid, "Ai-synthesized faces are indistinguishable from real faces and more trustworthy," *Proceedings of the National Academy of Sciences*, vol. 119, no. 8, pp. e2120481119, 2022.

[2] Zebin You, Xinyu Zhang, Hanzhong Guo, Jingdong Wang, and Chongxuan Li, "Are images indistinguishable to humans also indistinguishable to classifiers?," 2024.

[3] Yisroel Mirsky and Wenke Lee, "The creation and detection of deepfakes: A survey," *CoRR*, vol. abs/2004.11138, 2020.

[4] S. Lyu and H. Farid, "How realistic is photorealistic?," .

[5] Ruoyu Wu, Xiaolong Li, and Yang bin, "Identifying computer generated graphics via histogram features," .

[6] Andrew Gallagher and Tsuhan Chen, "Image authentication by detecting traces of demosaicing," .

[7] Fakhar Abbas and Araz Taeihagh, "Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence," .

[8] Shreyan Ganguly, Sk Mohiuddin, Samir Malakar, Erik Cuevas, and Ram Sarkar, "Visual attention-based deepfake video forgery detection," .

[9] Yogesh Patel, Sudeep Tanwar, Pronaya Bhattacharya, Rajesh Gupta, Turki Alsuwian, Inno Davidson, and ThokoZile Mazibuko, "An improved dense cnn architecture for deepfake image detection," .

[10] Xin Yang, Yuezun Li, and Siwei Lyu, "Exposing deep fakes using inconsistent head poses," .

[11] Ali Khodabakhsh, Raghavendra Ramachandra, Kiran Raja, Pankaj Wasnik, and Christoph Busch, "Fake face detection methods: Can they be generalized?," .

[12] Badhrinarayan Malolan, Ankit Parekh, and Faruk Kazi, "Explainable deep-fake detection using visual interpretability methods," in *2020 3rd International Conference on Information and Computer Technologies (ICICT)*, 2020, pp. 289–293.

[13] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," 2018.

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of stylegan," 2019.

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

[16] Google, "Mediapipe image segmenter," https://ai.google.dev/edge/mediapipe/, 2024, Accessed: 2024-12-21.

[17] Leland McInnes, John Healy, and James Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2020.

[18] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang, "Identification of deep network generated images using disparities in color components," .

[19] Sander Dieleman, "Diffusion is spectral autoregression," https://sander.ai/2024/09/02/spectral-autoregression.html.

[20] Tianqi Chen and Carlos Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Aug. 2016, KDD '16, p. 785–794, ACM.

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 8110–8119.

9. APPENDIX

9.1. Augmentations

1. Random brightness and contrast adjustment. Images undergo random alterations in brightness and contrast with a probability of 50%. This simulates varying lighting conditions, enabling the model to maintain performance across different illumination scenarios.

2. Gaussian blur. Gaussian blurring is applied to images with a 20% probability. This helps the model become invariant to slight out-of-focus conditions, improving its ability to generalize to real-world images that may not always be perfectly sharp.

3. Hue, saturation, and value shifting. The hue, saturation, and value (HSV) channels of images are randomly modified with a 30% probability. This accounts for color variability in different environments, ensuring the model can handle diverse color distributions.

4. Image compression. Images are compressed using JPEG compression with quality levels ranging between 75 and 90, applied with a 70% probability. This introduces compression artifacts that mimic real-world image degradation, enhancing the model's resilience to such distortions.

5. RGB channel shifting. The red, green, and blue (RGB) channels of images are independently shifted within a limit of ±20 units with a 30% probability. This augmentation improves the model's ability to handle variations in color channels, making it more robust to color discrepancies.

6. Random resized crop. A random resized cropping operation is performed on images to a fixed size of 224×224 pixels, with a scaling factor between 80% and 100%, applied with 100% probability. This ensures that the model learns to recognize objects at different scales and positions within the image frame.

7. Normalization. Image pixel values are normalized using the mean values (0.485, 0.456, 0.406) and standard deviations (0.229, 0.224, 0.225) for each respective channel. This standardization facilitates stable and efficient training of the neural network by maintaining consistent data distribution.

8. Noise injection. To enhance the model's resilience to noisy inputs, noise is introduced by randomly applying either Gaussian noise with a variance between 10.0 and 50.0 or ISO noise. Each type of noise is applied with a 50% probability, and the overall noise injection occurs with a 30% probability.

9. Tensor conversion. As the final step, augmented images are converted into tensor format. This conversion ensures compatibility with the neural network architecture, enabling efficient computation and seamless integration into the training pipeline.

9.2. Implementation Details

All experiments were conducted using T4 GPU and implemented in Pytorch. The model was trained with BCELoss using Adam optimizer with lr 1e-4 and weight decay 1e-3.