This Person Does (Not?) Exist

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

Introduction

The rise of generative AI made it increasingly easy to generate pictures of fake humans. Can these be consistently distinguished from the real ones?

- **Dataset**: ≈ 40,000 images, approximately evenly split between real and fake. Real people are scraped from Flickr by NVIDIA, while the rest are generated by StyleGAN, StyleGAN2 and Stable Diffusion 1.2. To avoid bias, both original and background-removed versions of the dataset were analyzed.

Figure 1: Class balance and data source.

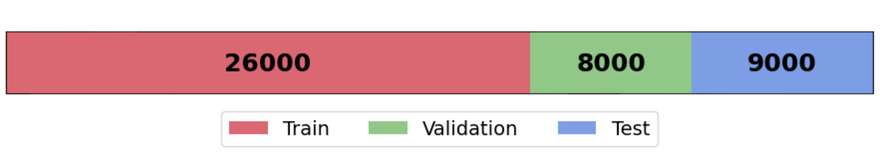

Figure 2: Train, validation and test split.

- **Images**: A study by Nightingale et al. [3] shows that images generated by StyleGAN2 are advanced enough to not be discernible by humans. A similar study by You et al. [4] was conducted for Stable Diffusion. To further emphasize the quality of the images, Figure 3 shows a general non-linear dimensionality reduction of the images.

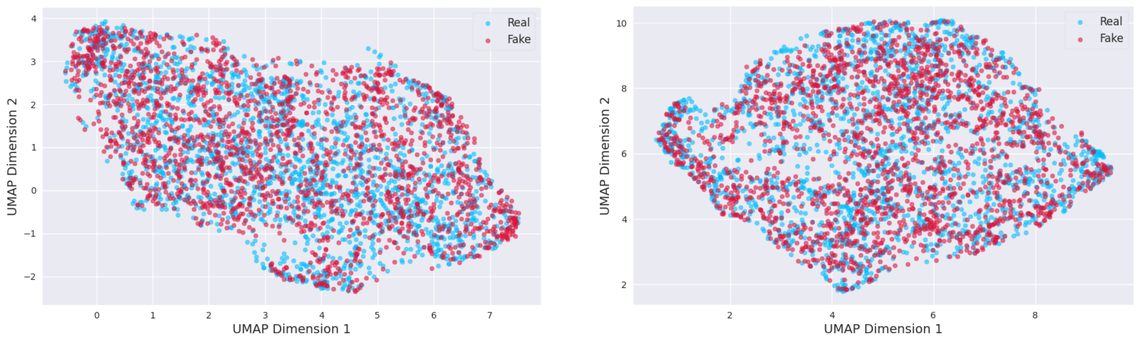

Figure 3: UMAP low-dimensional graph. On the left, with background, on the right, without.

EDA

- **Superimposion**: To investigate the possibility of recurring details that could be used as shortcuts, the faces were superimposed and inspected visually. Figure 4 shows noticeable differences in the general distribution, but no strong recurrent features.

Figure 4: Superimposion of fake and real faces, respectively.

- **Color distribution**: A study by Li et al. [2] suggests the predictive power of color distributions in identifying generated images. Thus, we computed the distributions per channel and assessed the significance of the difference using a K-S test.
- **Wavelet Energy**: According to a blog post by Dieleman [1], some differences might appear on the frequency domain. To consider this, we computed the wavelet energy at each decomposition level for both real and fake images, and performed an independent two-sample t-tests to assess the difference.

  Table 1 presents the resulting p-values of both computations, averaged per channel (color) and per decomposition level (wavelet).

Dataset	Histogram	Wavelet
Without Background	$< \mathbf{1 \times 10^{-8}}$	$\approx \mathbf{0}$
With Background	0.029	0.007

Table 1

Research Questions

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?

XGBoost Baseline

As a proof of concept and baseline, we developed an XGBoost model using only color distributions and wavelet energies (levels 1-3) as features, resulting in:

Dataset	Accuracy
Without Background	0.815
With Background	0.69

Modeling

The model combines features extracted by a pretrained EfficientNetV2 CNN with a MLP processing color distributions and wavelet energies. Aggressive data augmentation and regularization techniques like Dropout and L2 were applied to improve generalization and block dependency on specific features.

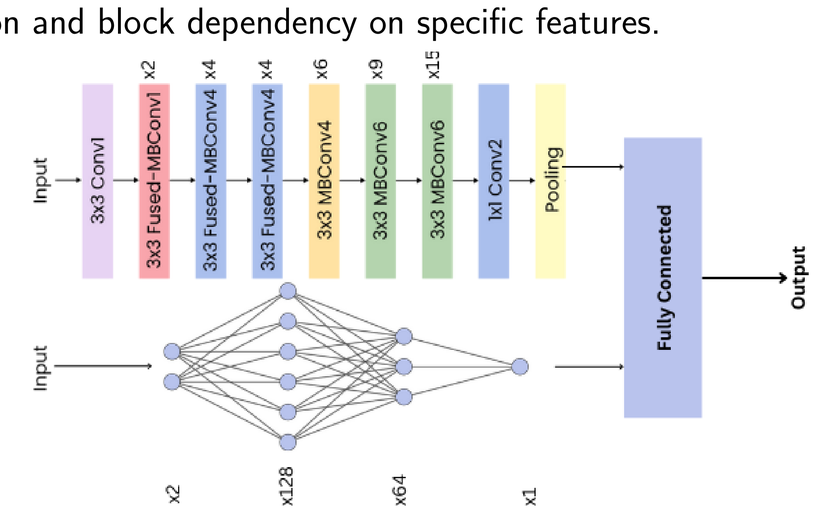

Figure 5: Architecture of the neural network.

Experimental Setup

- **Generalization:** To assess whether the model is able to generalize outside of the known distributions, OOD detection was performed on images generated from DALL-E 3, SD 2.0 and Midjourney.
- **Explainability:** To interpret the outputs of the model, SmoothGrad and GradCam were used to generate saliency maps of the predictions.

Results & Discussion

The model shows satisfying performances across all tests:

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

- **Similar Performance:** Similar performance between the two datasets suggests CNN-extracted features might carry most importance. This could also be a result of the aggressive augmentation, which forces the model to be less reliant on a specific feature, such as backgrounds.
- **OOD Performance:** The model shows impressive results in OOD images, with only one misclassification. This indicates suitability for a general-purpose detection task.
- **Saliency Maps:** The model shows understanding of higher-level features (e.g., eyes, mouths), with gradients focused on relevant areas (see Figure 6). Visual inspection is still challenging, as saliency maps generally provide limited information about the reasoning.

Figure 6: SmoothGrad saliency map and Grad-CAM visualization of selected fake images.

Future Work and Limitations

- **Future Work:** Explore unsupervised and self-supervised learning to capture complex features and use activation maximization to enhance model explainability.
- **Limitations:** The model is inherently temporary; releasing it could allow misuse as a GAN discriminator, making it unusable.

References

[1] S. Dieleman. Diffusion is spectral autoregression. https://sander.ai/2024/09/02/spectral-autoregression.html. [Accessed 07-12-2024].
[2] H. Li, B. Li, S. Tan, and J. Huang. Identification of deep network generated images using disparities in color components. doi: 10.1016/j.sigpro.2020.107616.
[3] S. J. Nightingale and H. Farid. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. doi: 10.1073/pnas.2120481119.
[4] Z. You, X. Zhang, H. Guo, J. Wang, and C. Li. Are image distributions indistinguishable to humans indistinguishable to classifiers? doi: 10.48550/arXiv.2405.18029.