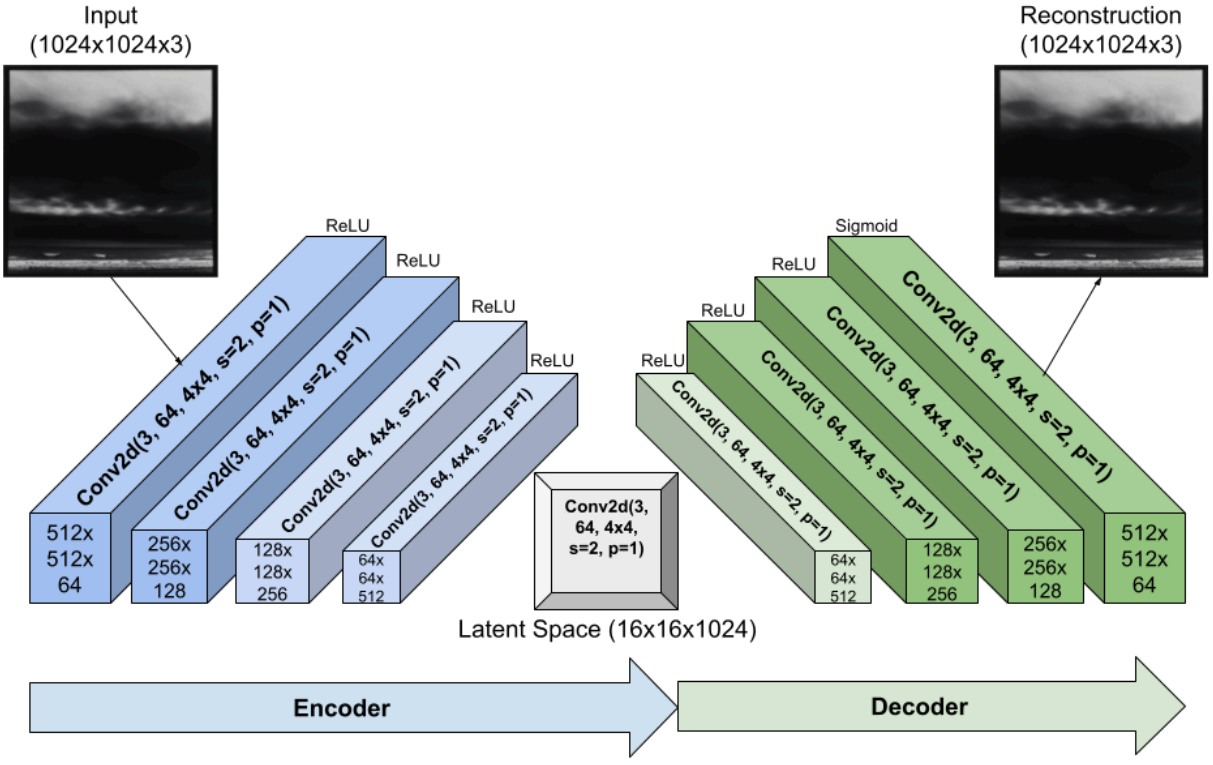
Alex Amari
2025

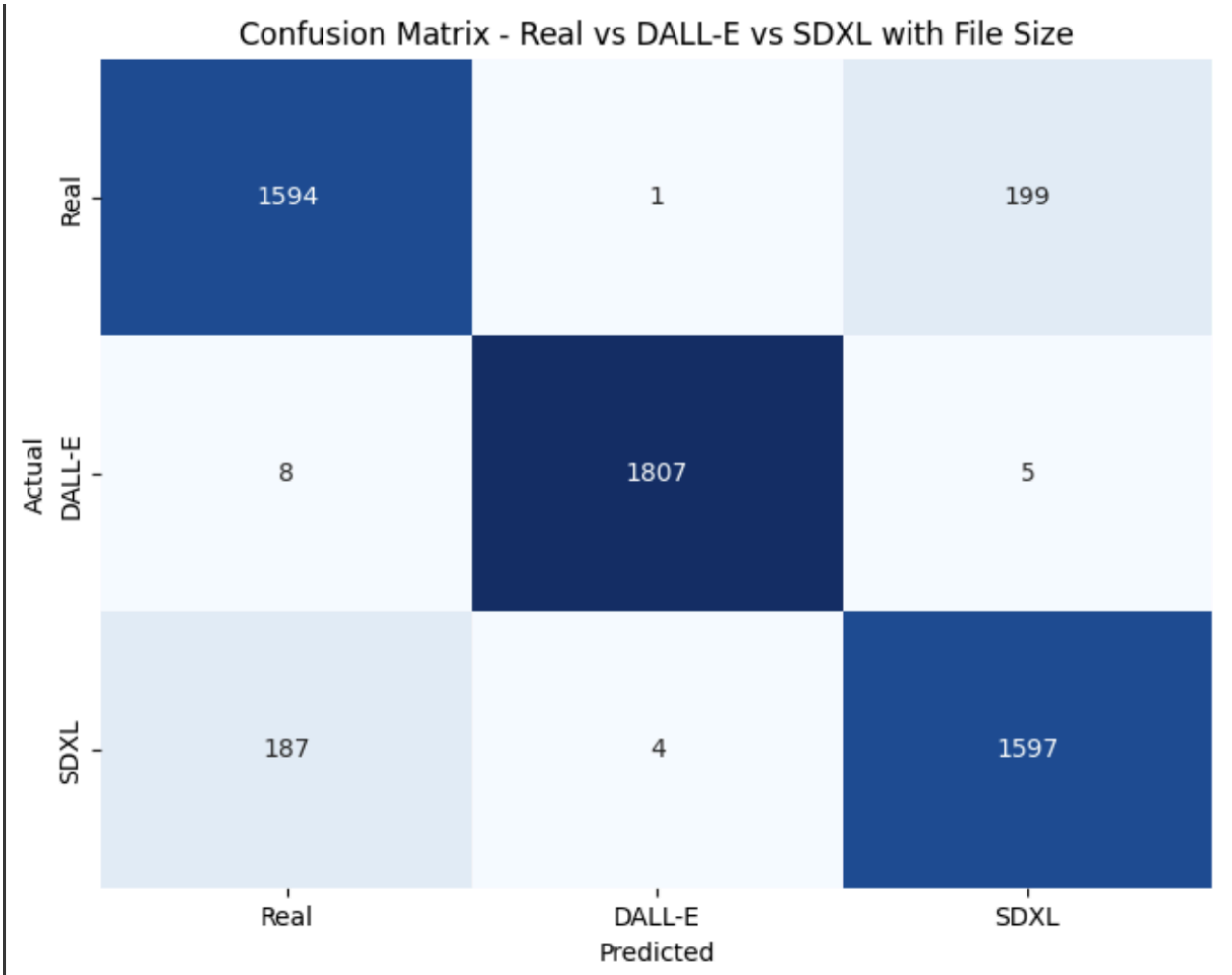
Intro

I've been training autoencoders to help detect "deep fakes" from various diffusion models. The idea is not only to discriminate between human and AI images, but also to identify the specific "model lineage" (if AI, which model made which image?). I used 10,000 real social media images from Instagram to generate diffusion "clones" by both Stable Diffusion XL and OpenAI's DALL-E, for a total of 30,000 images. I trained 3 different autoencoders, one on each image class (real, SDXL, DALL-E), and use their reconstruction errors and image compression size on a test set as features for a basic classifier (in principle, the lowest reconstruction error from an autoencoder should signal an image that is "native" to its corresponding model). Early results seem promising, but there's a ton more experimentation to be done on the architecture, more image models (particularly the new 4o-image from OpenAI, which is autoregressive as opposed to diffusion) file size regularization, and so on.

Sample Architecture



Some early results




```
Weighted F1 Score: 0.9252906237517631
Classification Report:
              precision    recall  f1-score   support

      DALL-E       0.89      0.89      0.89      1794
        Real       1.00      0.99      1.00      1820
        SDXL       0.89      0.89      0.89      1788

    accuracy                           0.93      5402
   macro avg       0.92      0.92      0.92      5402
weighted avg       0.93      0.93      0.93      5402
```