# detection of deep-fake profile images

goal: probability a image is generated by a given deep model

**BUT ALSO**

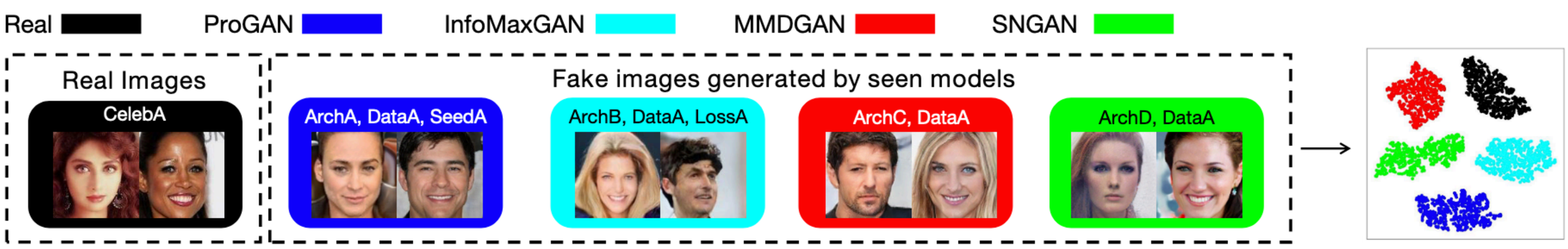
- mostly interested in fake **profile images**, not so interested in fake "kitchen" images
- interested on **social media** profile images


When uploading a picture on a social media platform, its quality gets degraded due to a jpeg compression. The jpeg "noise" makes the deep fake detection harder since generative models also leave a "noise" trace on the image.

***compressed** fake human faces*

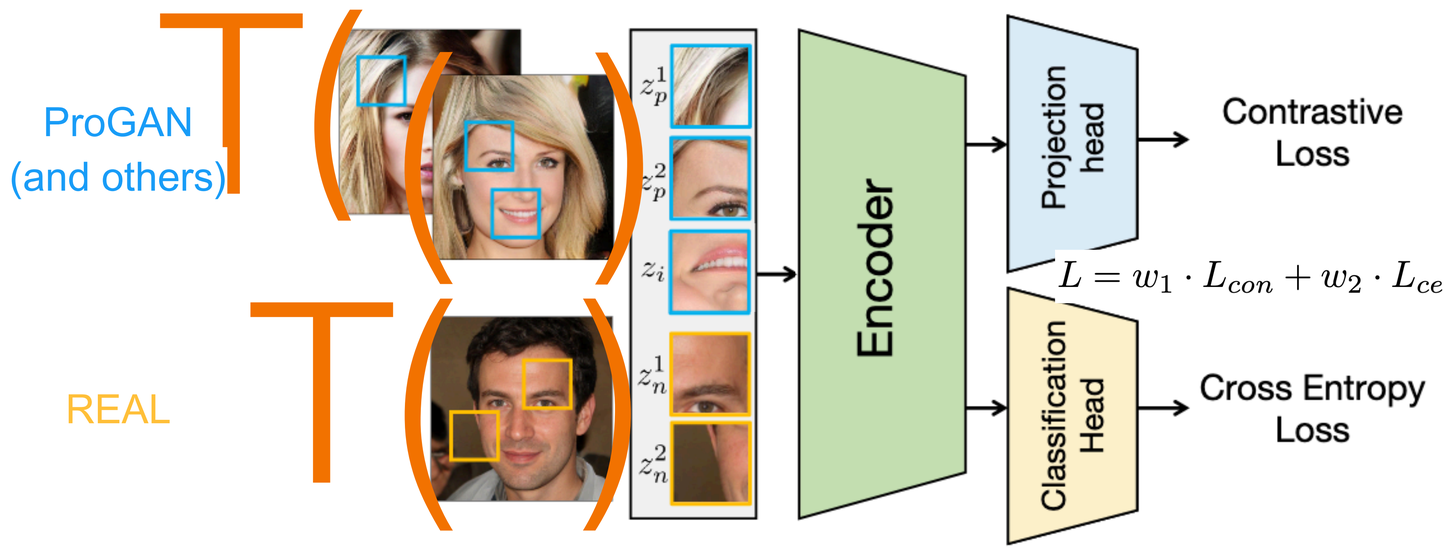
*Deepfake Network Architecture Attribution*
arXiv:2202.13843

robust architecture for fake-image classification

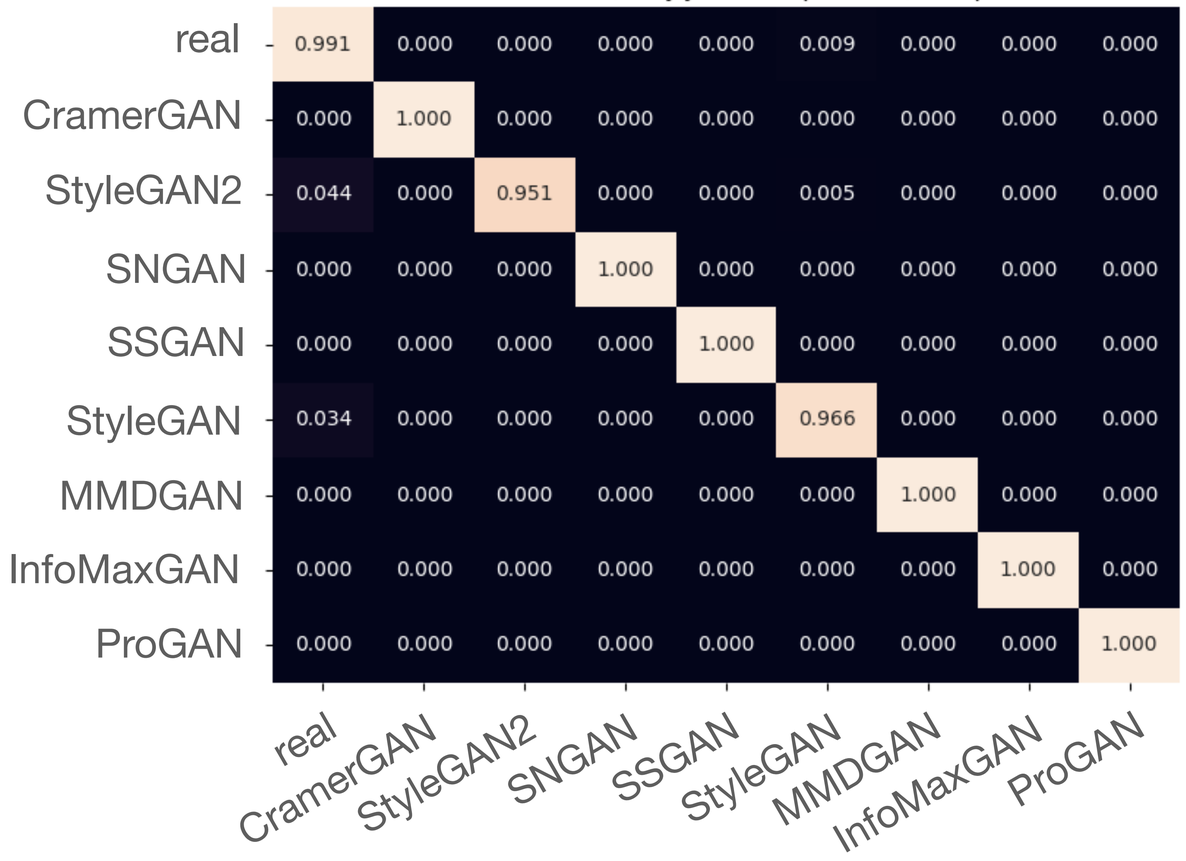



Images are divided in **patches** to force the model to recognise the **diffuse signature** left onto a generated image by GAN models.

On top of standard transformations, we added a **compression transformation** with multiple jpg-compression factors to the original paper contribution.




On validation set, only JPEG compression, 18 epochs




The original work was not able to detect fakes if the images were compressed. With the addition of compression transformations our GAN detection model works well on known GAN types, but would not be able to generalise to other types of image generation models (other GANs or diffusion models).

On the other hand, the model is robust for varying types of jpg compression.