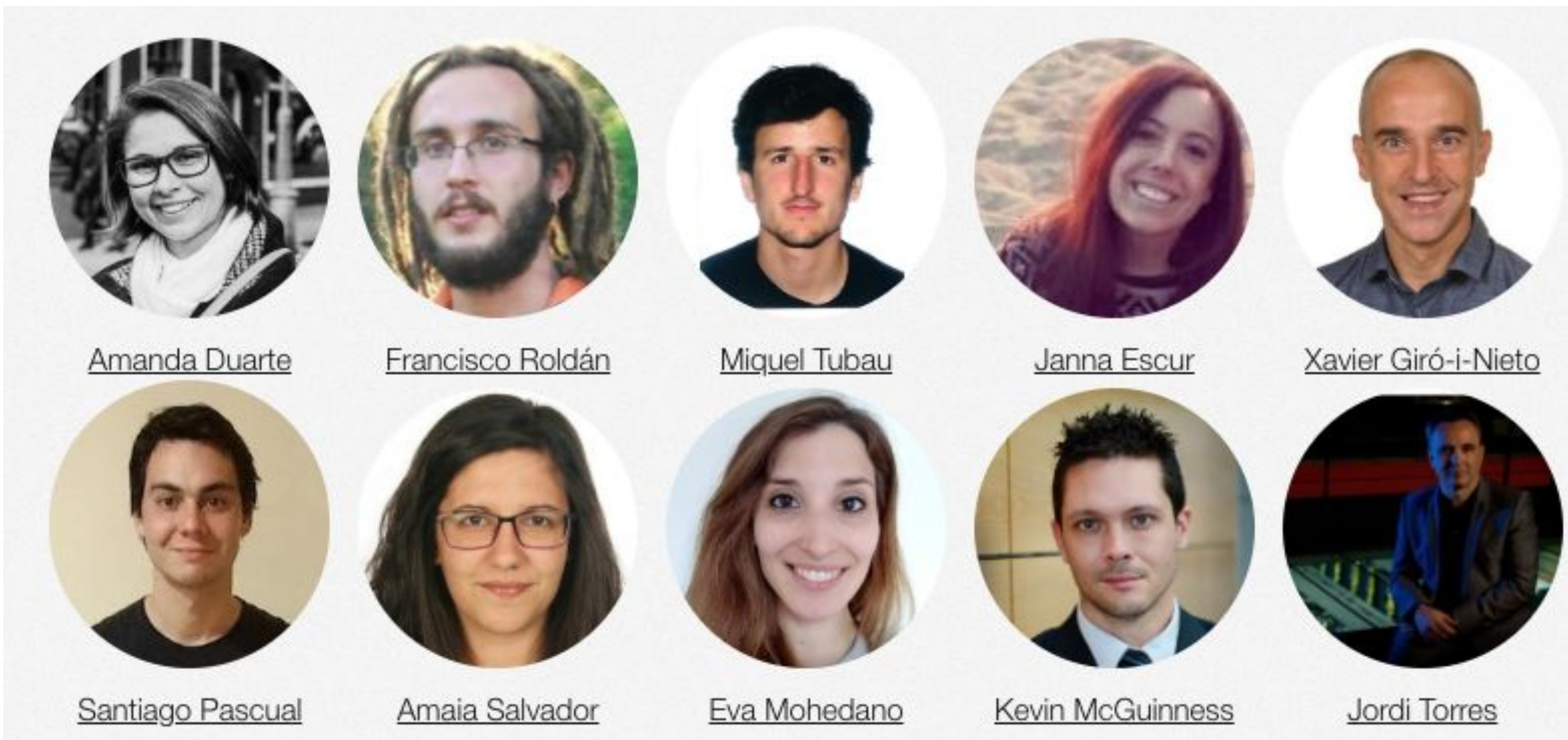
Wav2Pix

Speech-conditioned Face Generation using Generative Adversarial Networks

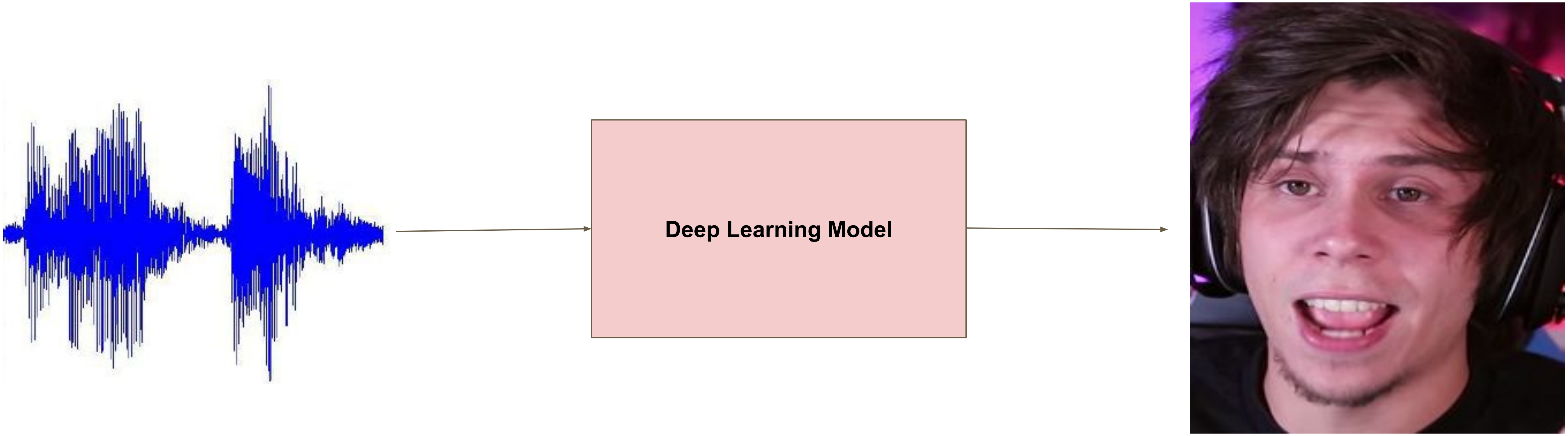


UPC

DCU

BSC

MOTIVATION



Speech Signal

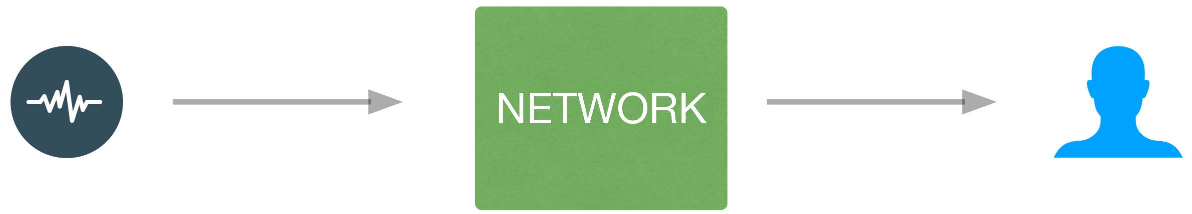
Face

MOTIVATION

- **Audio** and **visual** signals are the most common modalities used by humans to identify other humans and sense their emotional state

- Features extracted from these two signals are often **highly correlated**

- Roldán et. al. address this correlation proposing a face synthesis method using **exclusively** raw audio representation as inputs



Francisco Roldán Sánchez, "Speech-conditioned face generation with deep adversarial networks". Master in Computer Vision, 2018

RELATED WORK

GENERATIVE MODELS

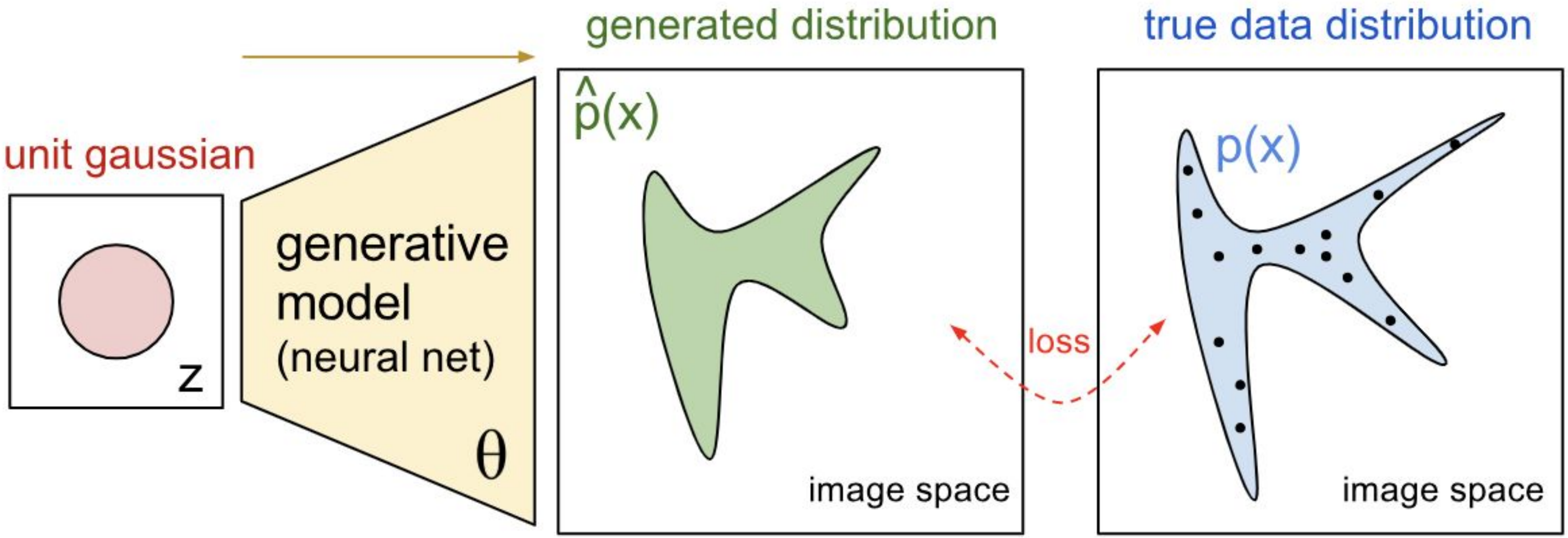
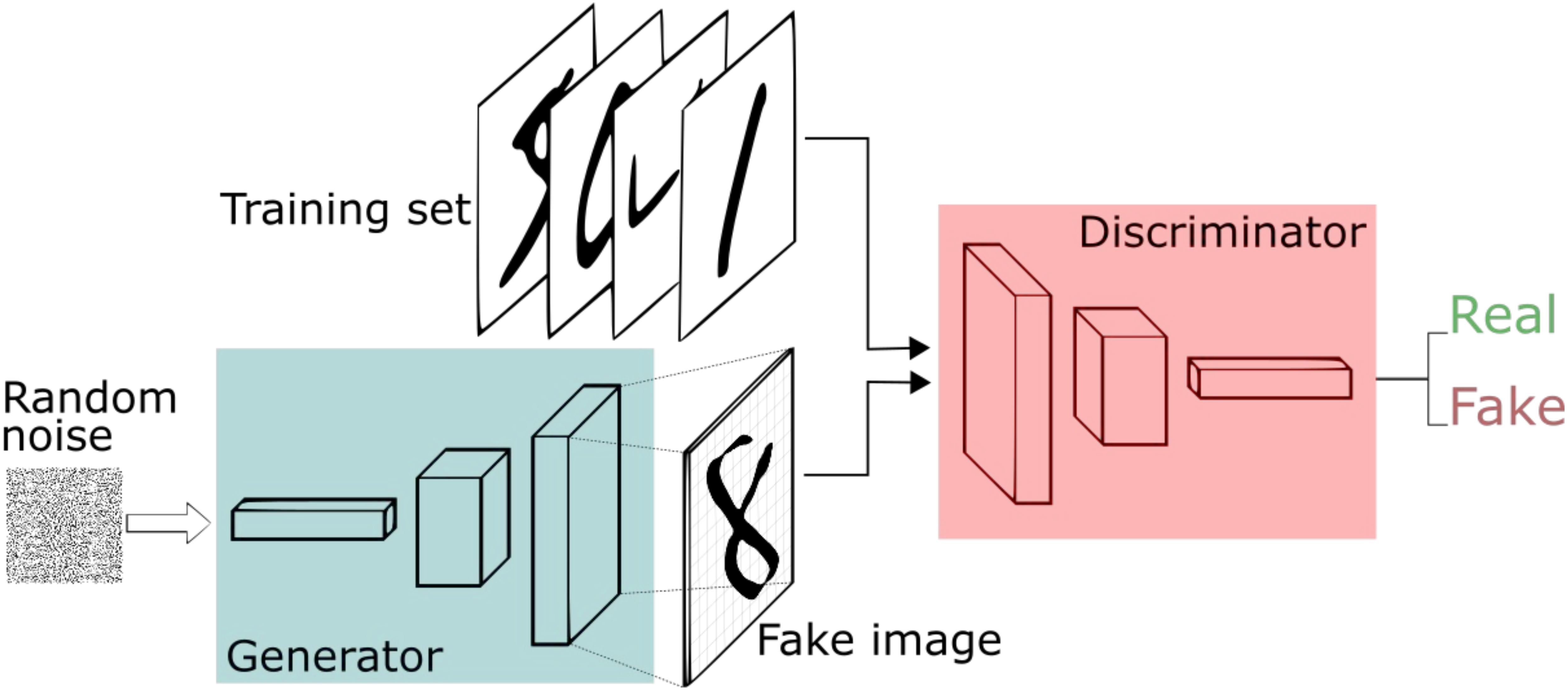


Image taken from https://blog.openai.com/generative-models/

GENERATIVE ADVERSARIAL NETWORKS



$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(\mathbf{x})}[\log \mathrm{D}(\mathbf{x})] + \mathbb{E}_{z\sim p_z(\mathbf{z})}[\log(1-\mathrm{D}(\mathrm{G}(\mathbf{z})))].$$

Goodfellow, Ian, et al. "Generative adversarial nets." *NIPS*. 2014.

GENERATIVE ADVERSARIAL NETWORKS

Imagine we have a counterfeiter (G) trying to make fake money, and the police (D) has to detect whether the money is **true** or **fake**.

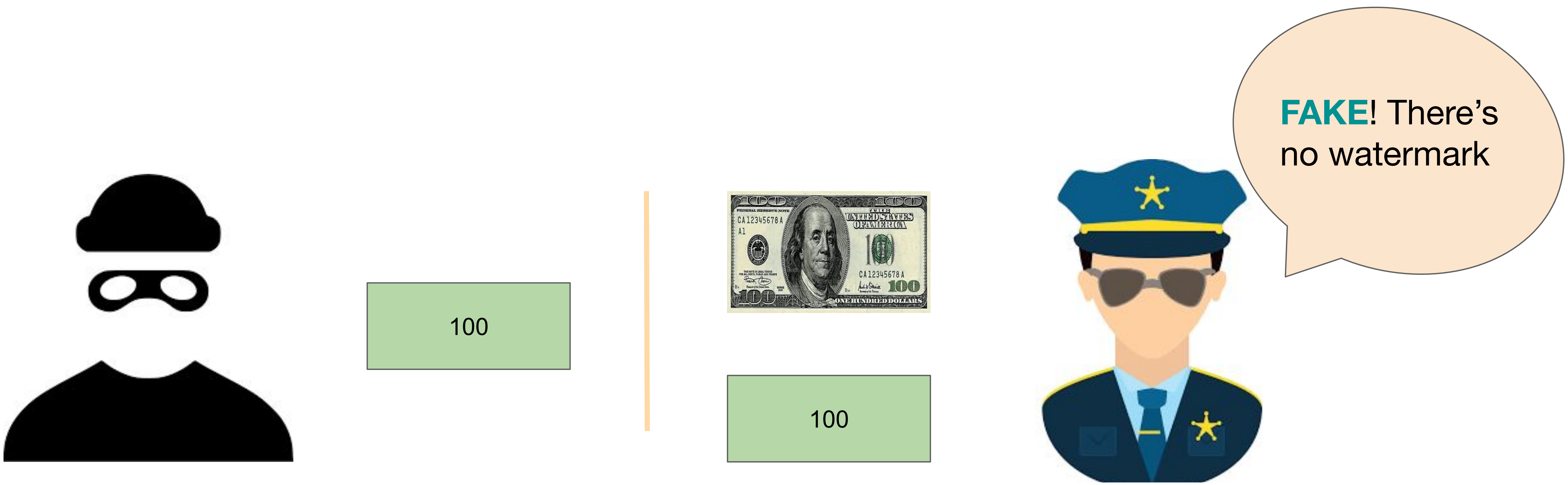
100

100

FAKE! It's not even green

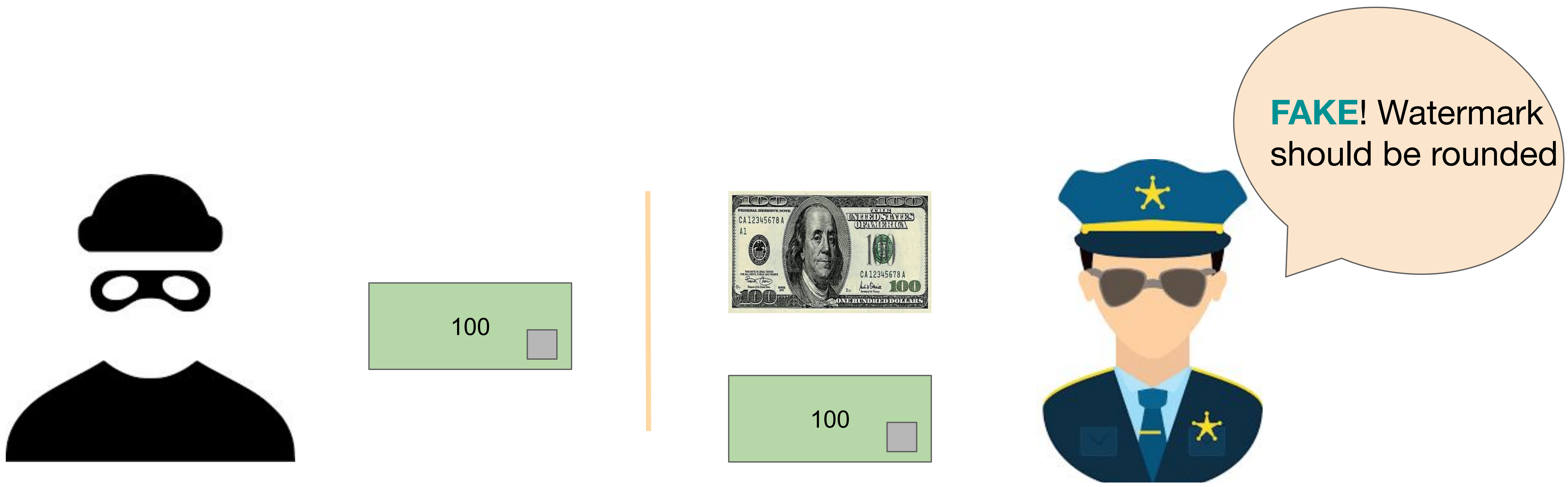
GENERATIVE ADVERSARIAL NETWORKS

Imagine we have a counterfeiter (G) trying to make fake money, and the police (D) has to detect whether the money is **true** or **fake**.



GENERATIVE ADVERSARIAL NETWORKS

Imagine we have a counterfeiter (G) trying to make fake money, and the police (D) has to detect whether the money is **true** or **fake**.



GENERATIVE ADVERSARIAL NETWORKS

After enough iterations:

Slide credit: Santi Pascual

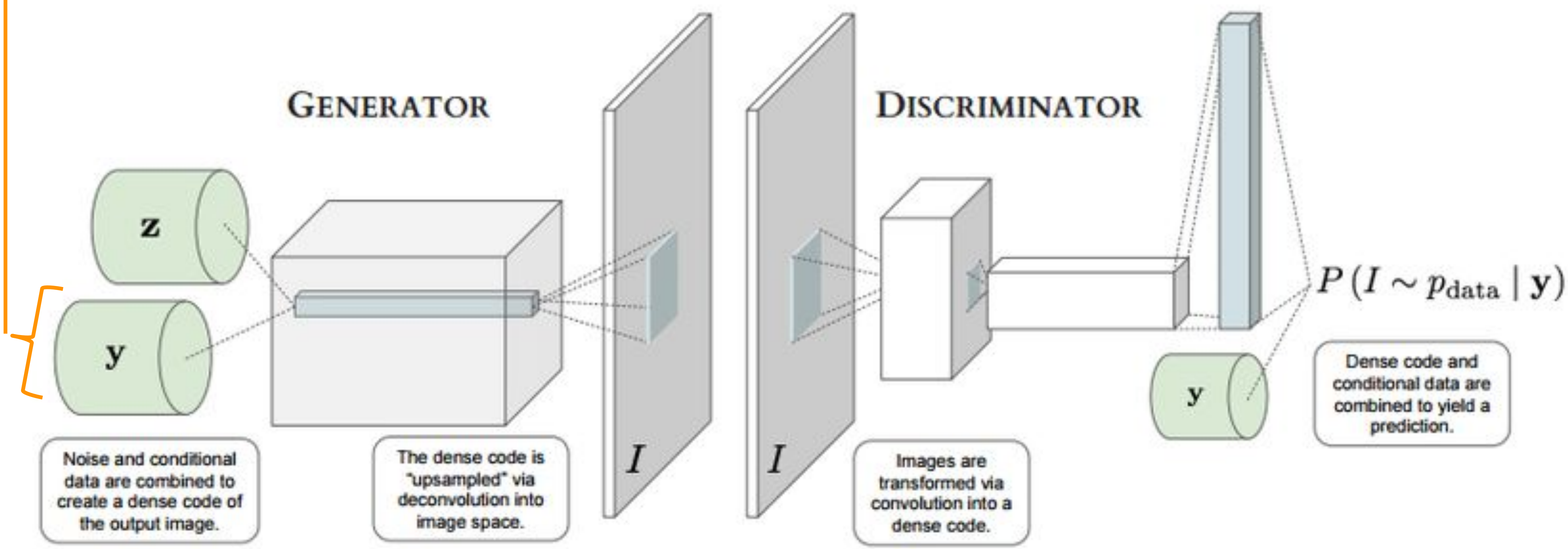
CONDITIONED GANs

In that case:

one-hot vector with the corresponding MNIST class

In our case:

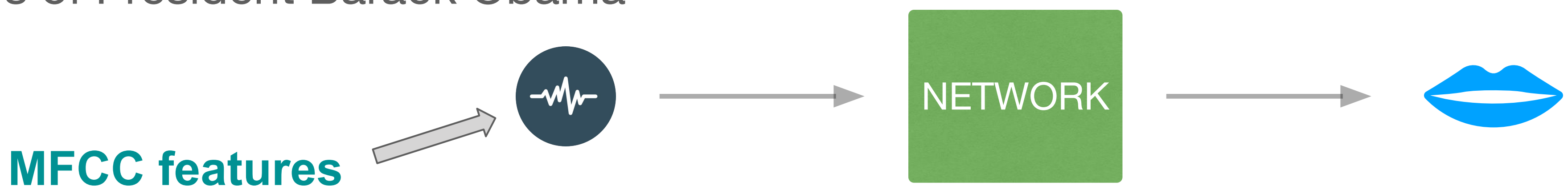
speech embedded in a 128 dimensional vector



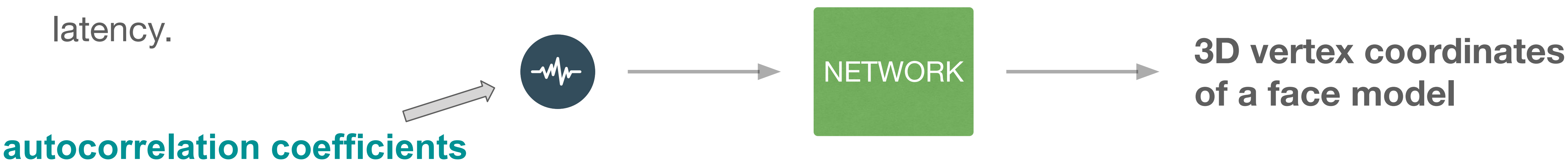
Mirza, Mehdi and Simon Osindero. "Conditional generative adversarial nets." NIPS (2014).

SPEECH-CONDITIONED IMAGE SYNTHESIS

- **Suwajanakorn** et. al. focused on animating a point-based lip model to later synthesize high quality videos of President Barack Obama



- **Karras** et. al. propose a model for driving 3D facial animation by audio input in real time and with low latency.



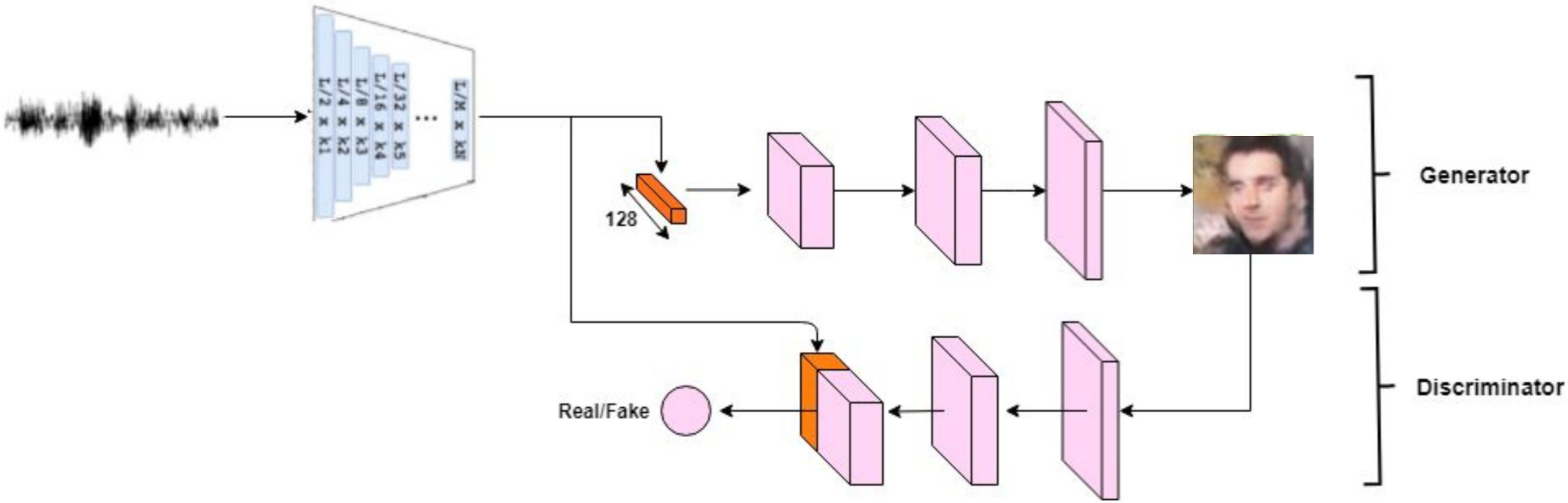
- **Chung** et. al. presented a method for generating a video of a talking face starting from audio features and an image of him/her (identity)

Chung, Joon Son, Amir Jamaludin, and Andrew Zisserman. "You said that?." BMVC 2017.
Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," ACM TOG, 2017.
Tero Karras,Timo Aila,Samuli Laine,Antti Herva, and Jaakko Lehtinen,"Audio-driven facial animation by joint end-to-end learning of pose and emotion," ACM TOG, 2017.

SPEECH-CONDITIONED IMAGE SYNTHESIS

- **Roldán** et. al. propose a deep neural network that is trained from scratch in an **end-to-end** fashion, generating a face **directly from the raw speech waveform without any additional identity information** (e.g reference image or one-hot encoding)

SPEECH-CONDITIONED FACE GENERATION WITH DEEP GANs

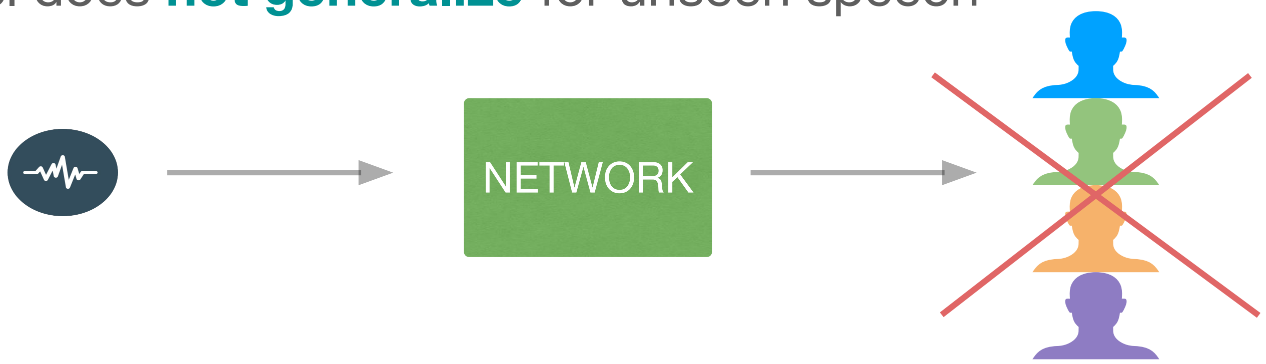


LSGAN **64x64 resolution** **Dropout instead of noise input**

Francisco Roldán Sánchez, "Speech-conditioned face generation with deep adversarial networks"
Santiago Pascual, Antonio Bonafonte, and Joan Serrà, "Segan: Speech enhancement generative adversarial network," Interspeech, 2017

SPEECH-CONDITIONED FACE GENERATION WITH DEEP GANs

- Roldán et. al. model does **not generalize** for unseen speech



- Inception Score metric used by Roldán et. al. evaluates the images in terms of quality but **not** in terms of **realism**

In this Project

Enhancement

- Search the optimal input audio **length**
- Add an **audio classifier** on top of the embedding
- Augment the capacity to generate **128x128** resolution

Evaluation

- Compute the accuracy of a **fine-tuned** VGG classifier
- Compute a novel approach based on a **face detection** system
- Perform an online **survey** assessed by humans

Francisco Roldán Sánchez, "Speech-conditioned face generation with deep adversarial networks"

DATASET

PREVIOUS DATASET

Good recording hardware

High amount of **frontal faces**

Wide range of **emotions**

youtubers_v1

Sex	Speakers	Faces	Speech (sec)
Male	29	26299	105196
Female	33	15900	63600
TOTAL	62	42199	168796

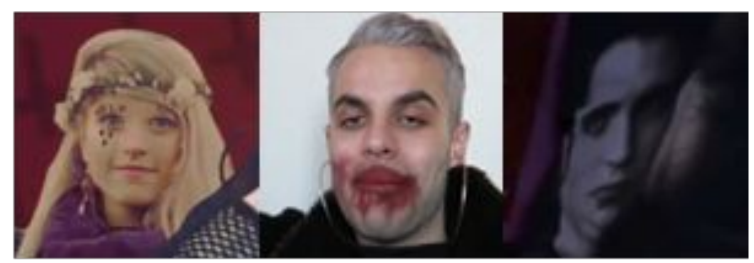
PREVIOUS DATASET

drawbacks

- **Imbalanced** dataset. Among the 62 youtubers, the amount of images/audios vary between 2669 and 52 pairs

- Notable amount of **false positives**

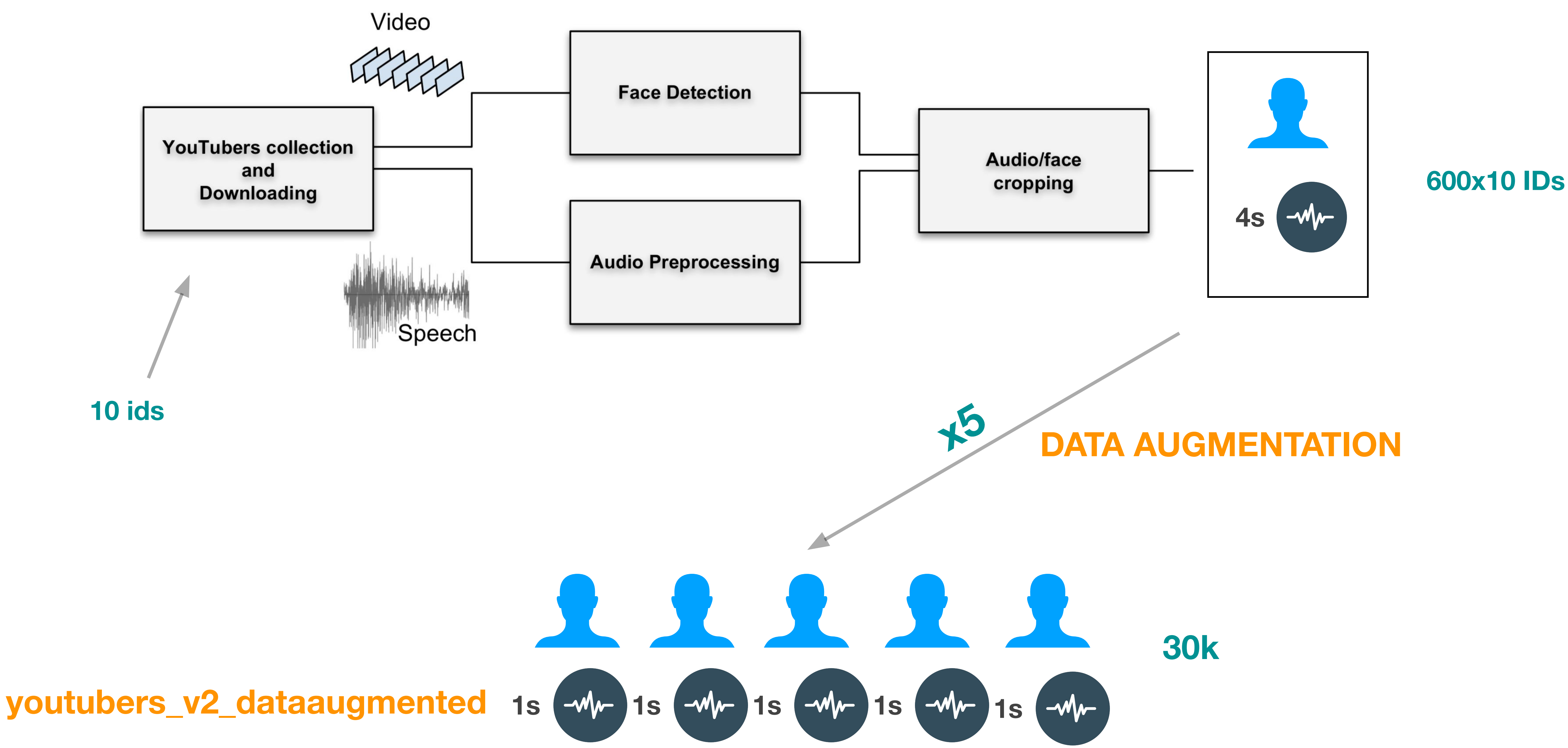
true identity



false positives

- Most of the speech frames were **noisy**
  - Background music in a post-process edition
  - Voice of a third person

youtubers_v2 - new dataset collection



DATASET

Roldán | Ours

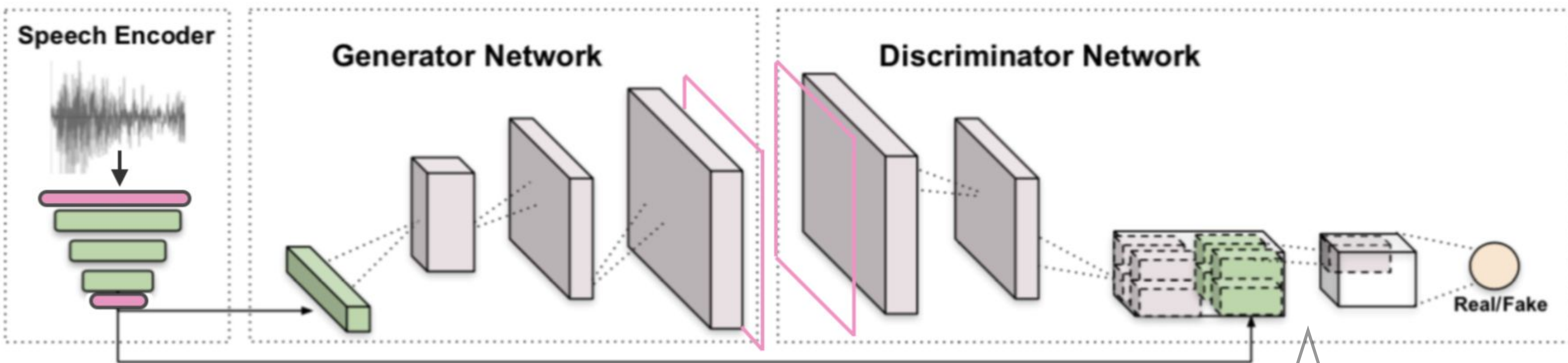
Features	youtubers_v1	youtubers_v2	youtubers_v2 data_augmented
Males	29	5	5
Females	33	5	5
Audio-face pairs	42199	6000	30000
Average audio-face pairs / ID	694	600	3000
Std audio-face pairs / ID	616	0	0
Audio duration (s)	4	4	1
Videos processed / ID	15	4	4
Balanced	False	True	True
Cleaned	False	True	True
Size in memory (GB)	7.4	1.8	2.1

ARCHITECTURE

Wav2Pix

ARCHITECTURE

end-to-end



segan
decoder
+
classifier

128

64 x 64 | 128 x 128

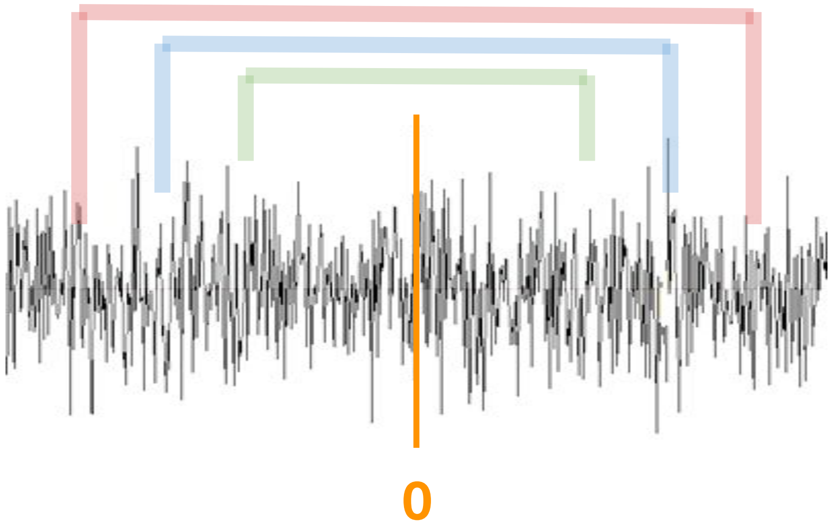
D loss
- real image loss
- fake image loss
- wrong image loss

G loss
- fake image loss
- softmax loss
- customized regularizers

ARCHITECTURE

contributions

- **Audio segmentation** module



- Speech **classifier**
    - 1-hidden NN with 10 output units

- **Additional** convolutional and deconvolutional layers
    - Kernel size: 4
    - Stride: 2
    - Padding: 1

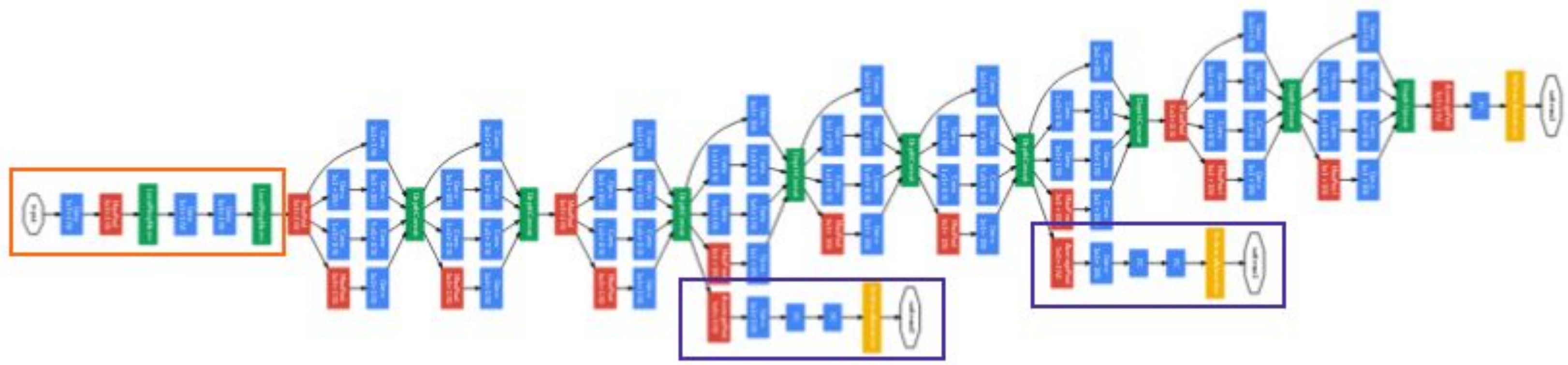
EVALUATION

EVALUATION

Fréchet Inception Distance

$$FID(r,g) = \|(\mu_r - \mu_g)\|_2^2 + \mathrm{Tr}(\sum_r + \sum_g - 2(\sum_r \sum_g)^{\frac{1}{2}})$$

- **Inception-v3** network pre-trained on **ImageNet**



- Results **not consistent** with human judgements
- **Little** amount of data to obtain reliable results
- The measure relies on an ImageNet-pretrained inception network, **far from ideal** for datasets like **faces**

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in Advances in Neural Information Processing Systems, 2017, pp. 6626– 6637.

EVALUATION

VGGFace fine-tuned classifier

- Network proposed by the Visual Geometry Group department of Engineering Science (University of Oxford)

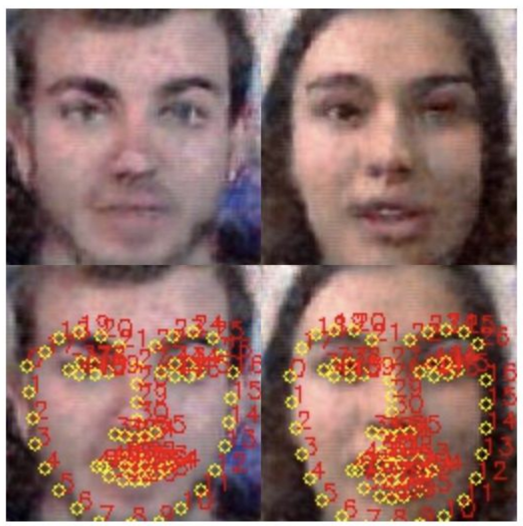
	Roldán	Ours
Real data	100	100
Generated data for seen speech	56.34	76.81
Generated data for unseen speech	16.69	50.08

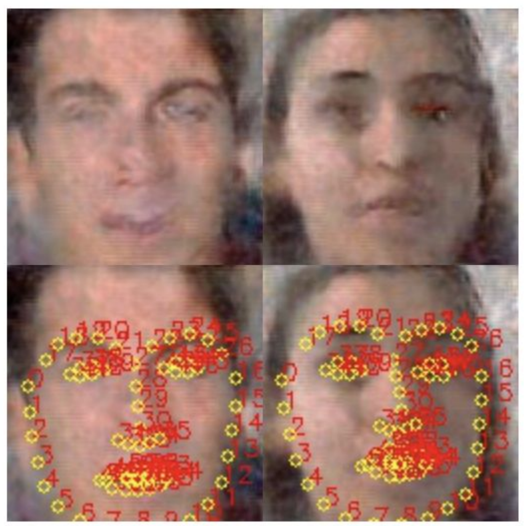
- **Improvement** of our model in preserving the identity
- Bearing in mind the metric is sensible to image quality, and the probability of confusion is 90%, the results are **promising**.

O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in British Machine Vision Conference, 2015.

Facial Landmark Detection ratio

- **Robustness** to image quality





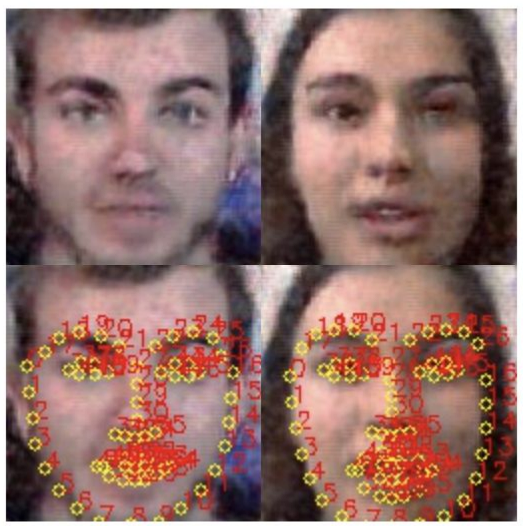
	Roldán	Ours
Real data	75.02	72.48
Generated data for seen speech	61.76	84.45
Generated data for unseen speech	60.81	90.25

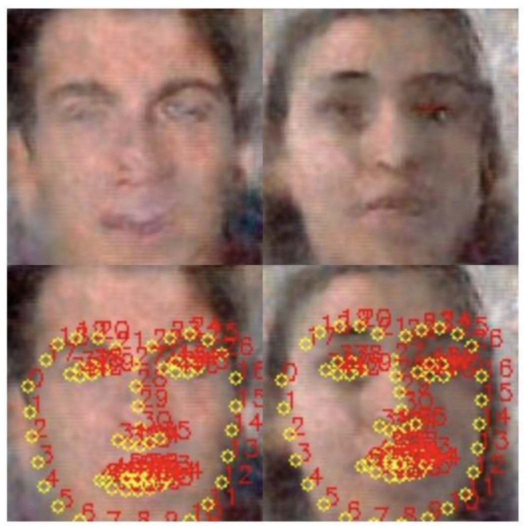
- **90%** of the generated images of our model for unseen speech can be considered as faces

EVALUATION

Facial Landmark Detection ratio

- **Robustness** to image quality





	Roldán	Ours
Real data	75.02	72.48
Generated data for seen speech	61.76	84.45
Generated data for unseen speech	60.81	90.25

- **90%** of the generated images of our model for unseen speech can be considered as faces

EVALUATION

Online survey

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N}$$

- **42** people have been asked to answer **2** questions for **32** different pairs of images:

Real Image (baseline)

Compare the quality of the generated image with respect to the real one (5-identical, 4-good-, 3-fair, 2-poor, 1-bad)

	1	2	3	4	5	
Bad	○	○	○	○	○	Identical

Generated image

Could you recognize the real person (appearing in the baseline image) from the generated image?

- ○ Yes
- ○ No
- ○ Not sure

MOS
2.09

% NOT SURE	% NO	% YES
14	52	34

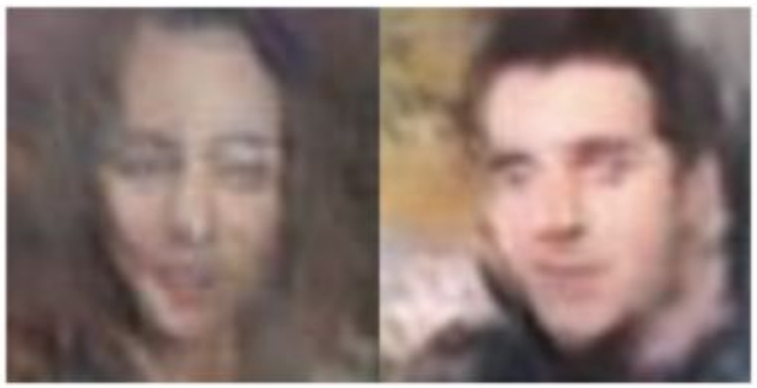
- Not reliable results
- This metric should be further improved

EXPERIMENTS

EXPERIMENTS

datasets comparison

Best quality images manually selected



youtubers_v1

youtubers_v2

youtubers_v2
data augmented

Facial landmark detection ratio (%)

youtubers_v1	youtubers_v2	youtubers_v2 Data Augmented
60.81	71.47	90.25

The following experiments have been performed with this dataset

EXPERIMENTS

input audio length

Best quality images manually selected w.r.t the audio length

0.3 s 0.7 s 1s

Fine-tuned VGGclassifier accuracy in % w.r.t the audio length (in seconds)

0.3	**0.7**	**1**
89.12	81.16	90.25

The following experiments have been performed with this length

EXPERIMENTS

input audio length

true identity: Jaime Altozano

true identity: Mely

The more **voice** frames in the audio, the easier for the network to learn the identity

EXPERIMENTS

image resolution

The following experiments have been performed with 128x128 image resolution

EXPERIMENTS

identity classifier

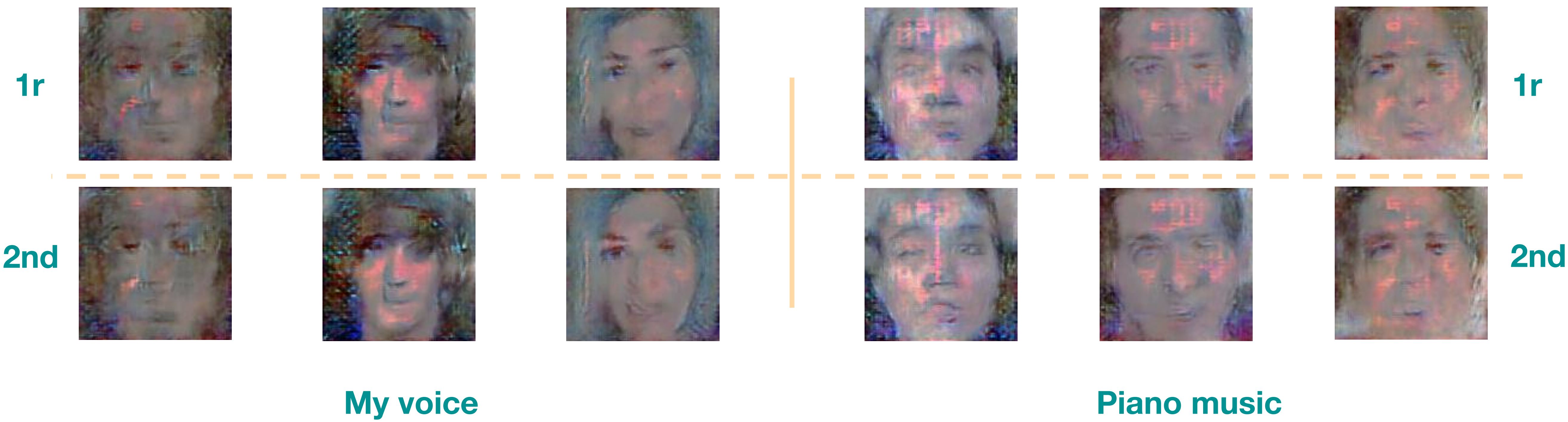
Fine-tuned VGGFace classifier accuracy in % w.r.t the model

Roldán	Ours
16.69	50.08

Close to ramdomness! (10%)

EXPERIMENTS

image generation for unseen voice



The network does **not generalize** for unseen IDs!!

EXPERIMENTS

audio interpolation

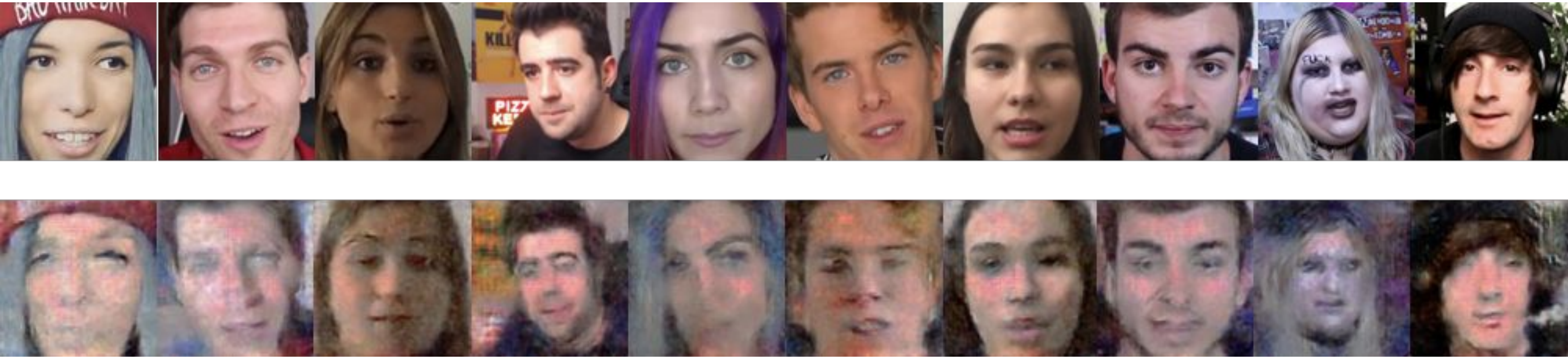
0.9×*audio_mely*+0.1×*audio_jaime*

0.4×*audio_mely*+0.6×*audio_jaime*

The network does not generate faces for audios which do not contain distinguishable voice. The model has learned to **identify speech** in audio frames

EXPERIMENTS

image generation for audio averages



The model performs a good **generalization** for **unseen** speech of **seen** IDs

CONCLUSIONS

CONCLUSIONS

In comparison to Roldán et. al. network, our contributions allows the final model:

- Generate images of **higher quality** due to the network's capacity increase ✓
- Generate **more face-looking** images for unseen speech ✓
- Preserve the **identity** better for unseen speech ✓
- Obtain better results with a **smaller dataset** (~70% smaller in terms of memory size) ✓
- Obtain results that can be evaluated in terms of quality, face appearance and identity preservation with three different metrics ✓

However,

- No generalization is achieved for **unseen** ID's ✗
- The dataset needs to be very clean in order to obtain notable results. The building process is very time-consuming ✗