Paper info

- **Title**: ICface: Interpretable and Controllable Face Reenactment Using GANs
- **Authors**: Soumya Tripathy, Juho Kannala and Esa Rahtu
- **Link**
- **Year**: 2019

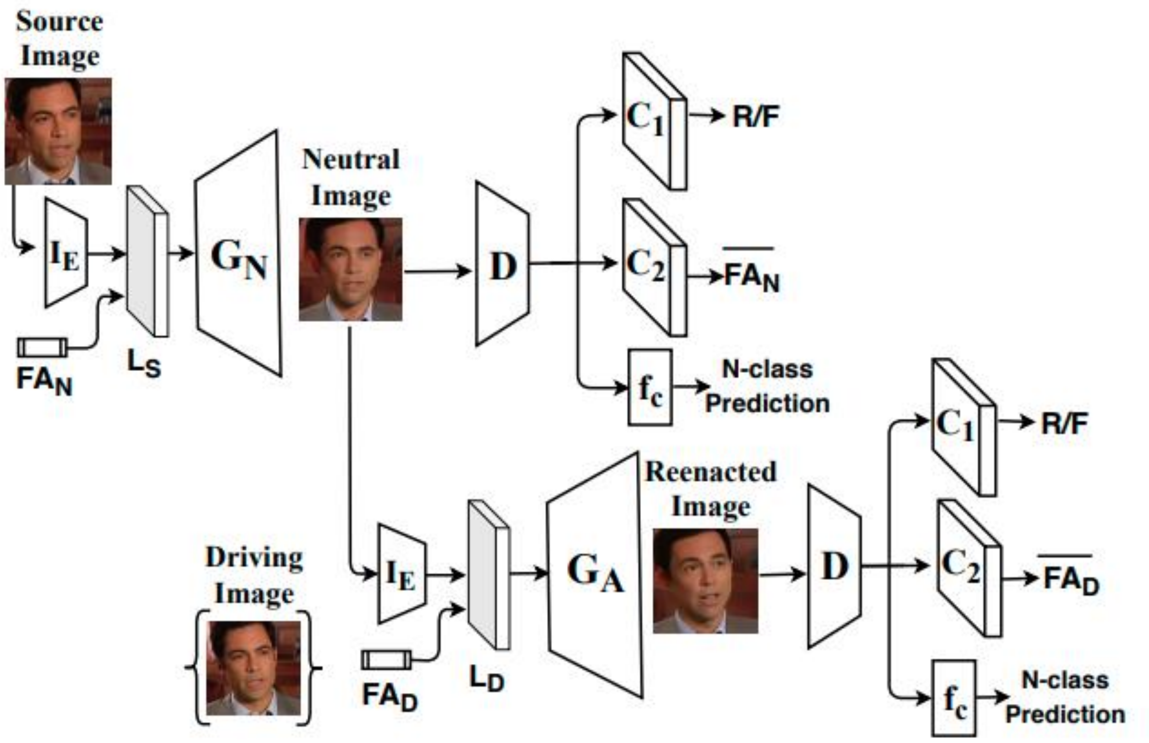
Task formulation & methods

Idea - face animator to controll facial expression and pose. Face control is done using Actual Units for expression and angles for pose. AUs and angles form together a feature tensor. An animator is implemented as a 2-stage NN, learned in self-supervised manner using a large video collection.

Method/pipeline description

- Main task - take image as a source and produce a face image depicting the source identity with the desired features (further driving features, or driving image).
- Traditional approach - fit detailed 3d face model on source img, and create animation of the model. Problem - requires too much effort.
- Recently - animation is directly formulated as an end-to-end learning problem, where the necessary model is obtained implicitly using a large data collection. Problem - lacks interpretability and does not allow selective editing.
- So authors proposed a GAN based system that is able to reenact realistic emotions and head poses for a wide range of source and driving identites. Allows selective editing and extensive control.

Architecture



The architecture of our model consists of four different subnetworks: **image encoder, face neutraliser, face generator**, and **discriminator**.

- **Image encoder I_E**

A network that maps maps the input face image into an equal size feature tensor. The network has a hourglass architecture consisting of convolutions and deconvolutions with normalization and activation layers.

- **Neutralizer G_N**

The neutralizer is a generator network that transforms the feature representation into representation with a neutral pose and facial expression. The architecture of the GN network consists of strided convolution, residual blocks and deconvolution layers. Inspired by CycleGAN arcitecture.

- **Generator G_A**

The generator network transforms the feature representation of the neutral face into the final reenacted output image. The output image is expected to be same as source identity with driving features applied. The architecture of the GA network is similar to that of GN .

- **Discriminator D**

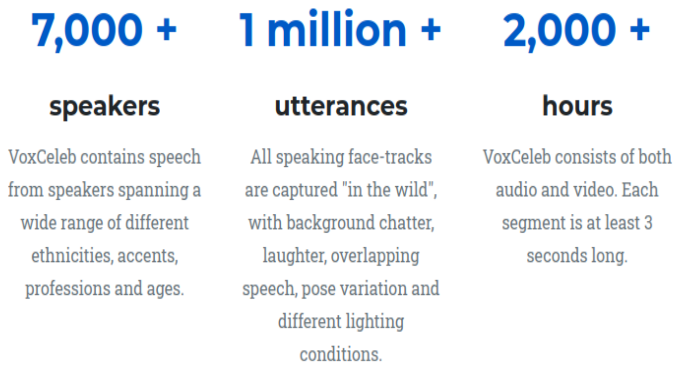
The discriminator network performs three tasks simultaneously:

- evaluates the realism of the neutral and reenacted images through C1
- predicts the facial attributes through C2
- classifies the identity of the generated face through FC layer with softmax

The blocks C1 and C2 consist of convolution block with sigmoid activation. The overall architecture consists of strided convolution and activation layers. The same discriminator with identical weights is used for GN and GA.

Training

- **dataset** - VoxCeleb, publically available. -->
- **process**: take one frame of the same video as source image, next as driving. Extract features from driving, and feed them to the network as driving features FA_D
- **losses** - a weighted combination of the following losses: **Facial attribute reconstruction loss**, **Identity classification loss**, **Reconstruction loss**.



7,000 +

speakers

VoxCeleb contains speech from speakers spanning a wide range of different ethnicities, accents, professions and ages.

1 million +

utterances

All speaking face-tracks are captured "in the wild", with background chatter, laughter, overlapping speech, pose variation and different lighting conditions.

2,000 +

hours

VoxCeleb consists of both audio and video. Each segment is at least 3 seconds long.

Results

The results and comparison are better seen from images in the paper, as face reenacment is a very visual field. Better see paper ilustrations.