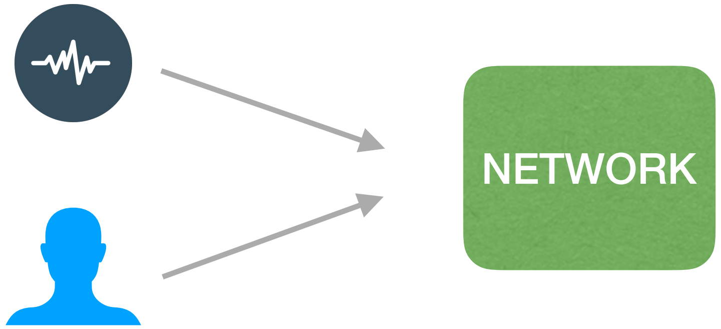
Wav2Pix

Speech-conditioned face generation using Generative Adversarial Networks

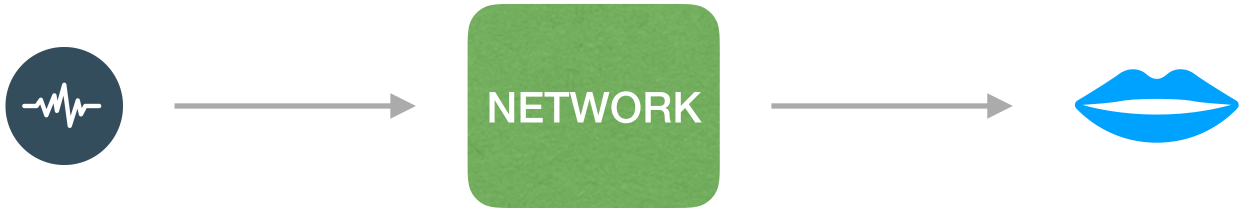
Amanda Duarte[†⋆], Francisco Roldan[⋆], Miquel Tubau[⋆], Janna Escur[⋆], Santiago Pascual[⋆] Amaia Salvador[⋆], [‡]Eva Mohedano, Kevin McGuinness [‡], Jordi Torres[†⋆], Xavier Giro-i-Nieto[†⋆]

MOTIVATION

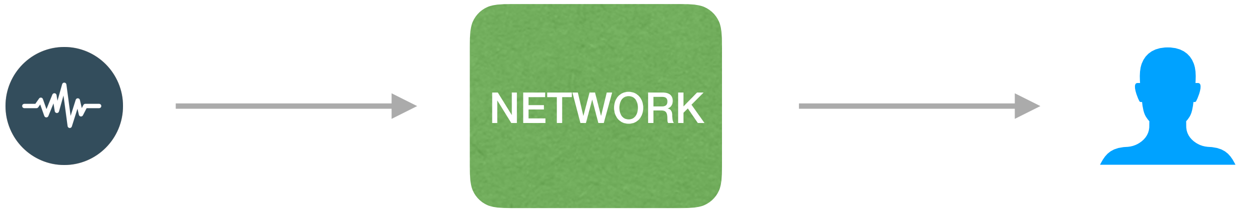
- Chung et al. presented a method for generating a video of a talking face starting from audio features and an image of him/her (identity)



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



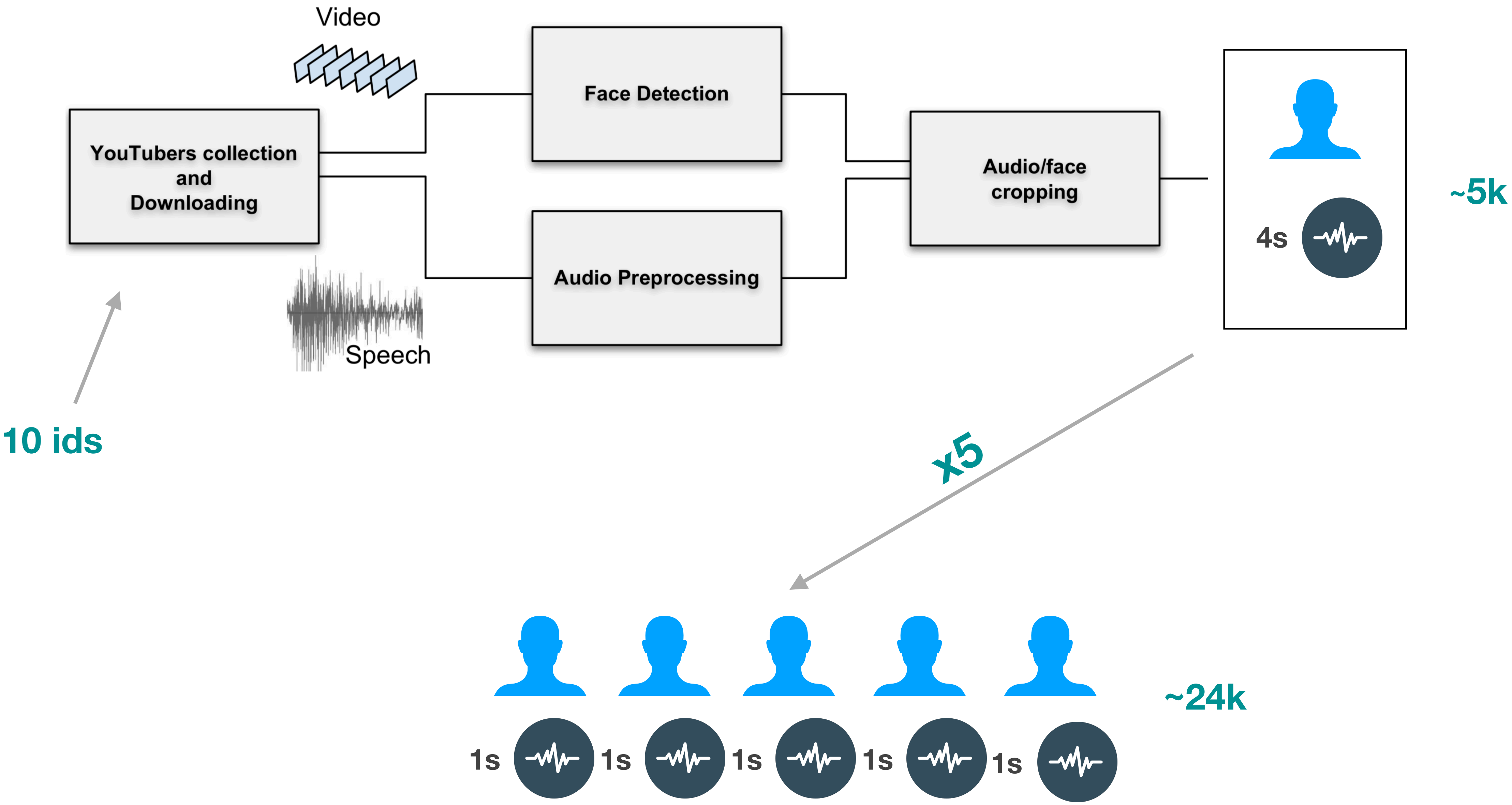
- **We** aim to generate the whole face image at **pixel level**, conditioning **only** on the raw speech signal (i.e. without the use of any handcrafted features) and without requiring any previous knowledge (e.g speaker image or face model).



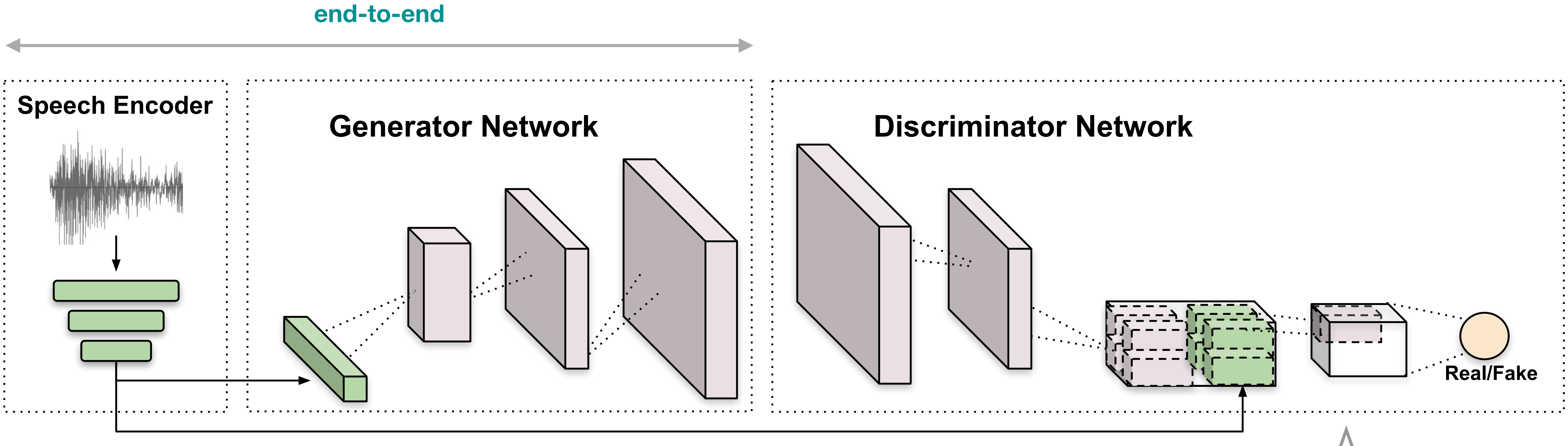
MOTIVATION

We 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)

DATA


Video
YouTubers collection and Downloading
Speech
10 ids
Face Detection
Audio Preprocessing
Audio/face cropping
4s
~5k
x5
1s
1s
1s
1s
1s
~24k

ARCHITECTURE



segan
decoder
+
classifier

128

64 x 64 | 128 x 128

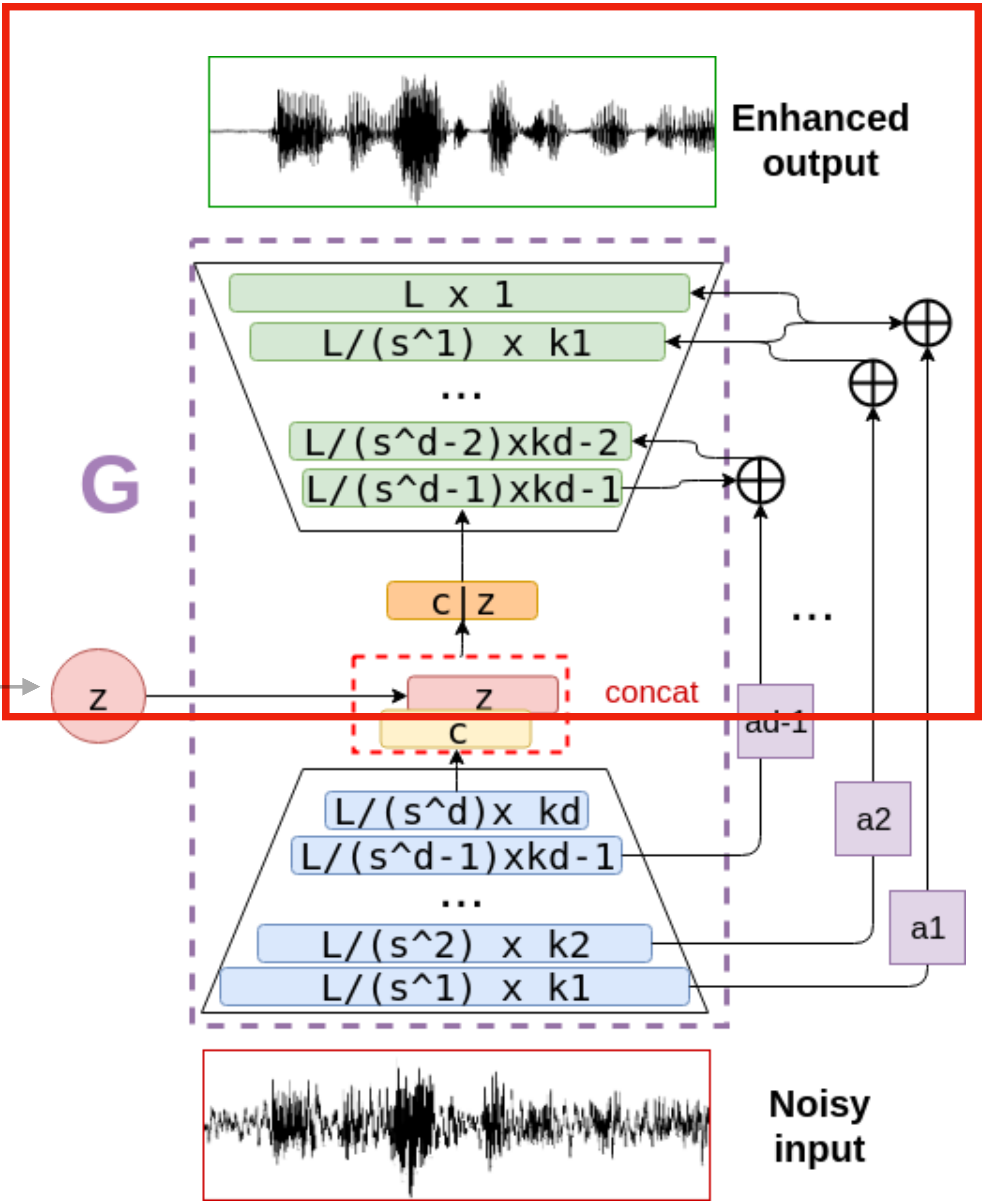
D loss

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

G loss

- fake image loss
- softmax loss
- customized regularizers

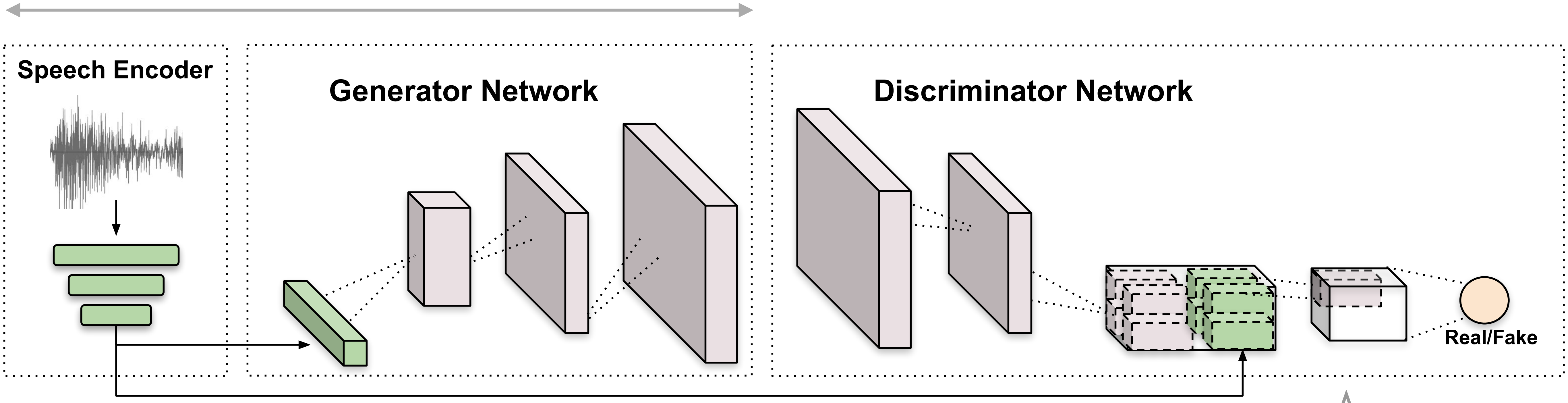
SEGAN


Enhanced output
G
L x 1
L/(s^1) x k1
...
L/(s^d-2)xkd-2
L/(s^d-1)xkd-1
c | z
...
z
concat
c
ad-1
L/(s^d)x kd
L/(s^d-1)xkd-1
...
a2
L/(s^2) x k2
a1
L/(s^1) x k1
Noisy input

128

ARCHITECTURE

end-to-end



segan
decoder
+
classifier

128

64 x 64 | 128 x 128

D loss

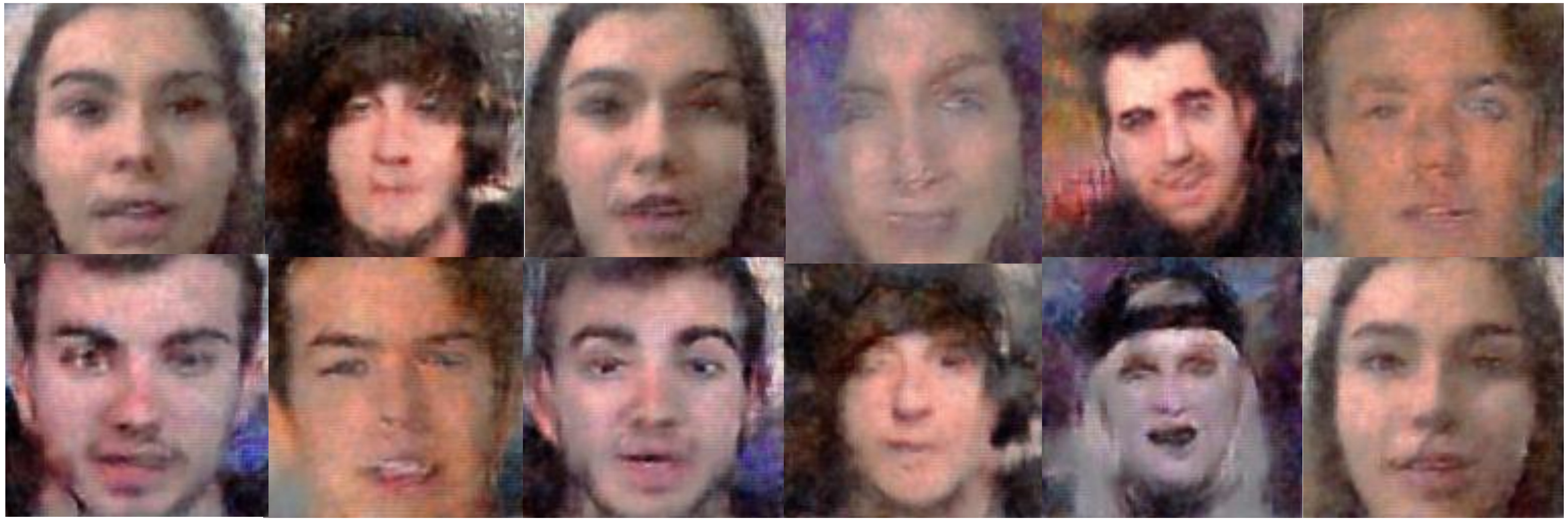
- real image loss
- fake image loss
- wrong image loss

G loss

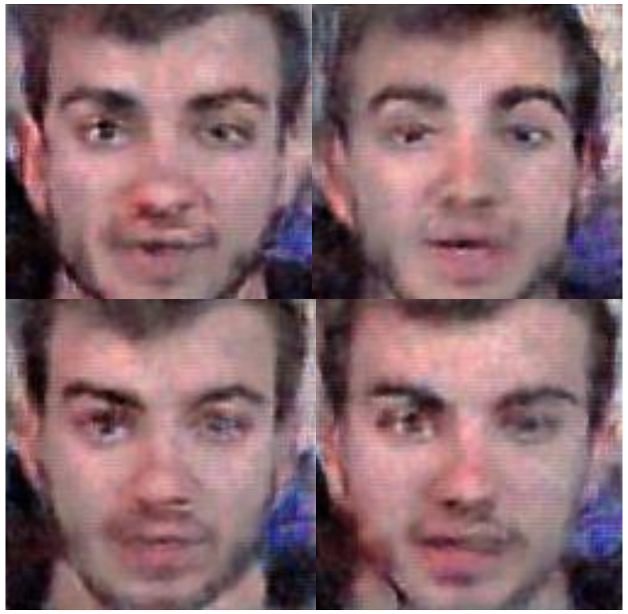
- fake image loss
- softmax loss
- customized regularizers

RESULTS

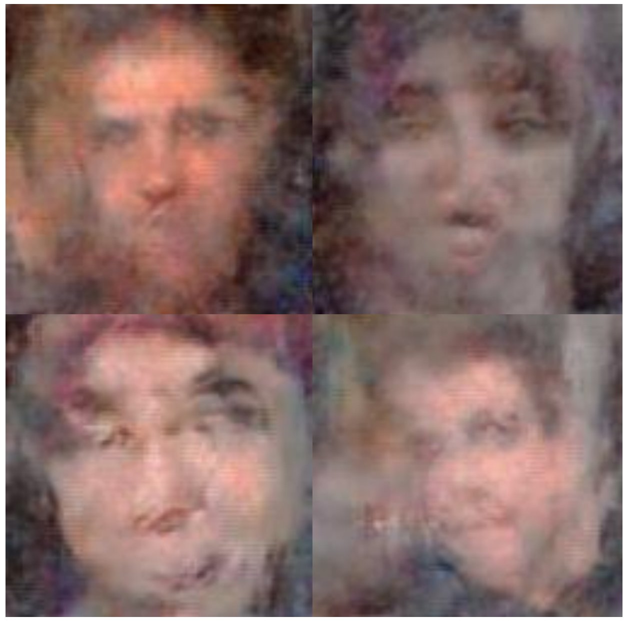
BEST RESULTS



EXPRESSION



BAD RESULTS



FURTHER STEPS

- Generate faces for **unseen** IDs and se if we obtain realistic images

- Evaluation metrics
  - **FID** distance did not work well. Our best model did not obtain the lowest score

- Focus on **expressivity**
  - Is the audio expression similar to the one seen in the generated image?
  - Include an emotion classifier in the pipeline

THANK YOU!