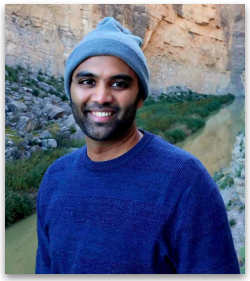
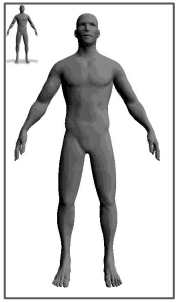
StyleWav: Guiding Image Synthesis Using Audio

Phuc (Jerry) Ngo[1], Swami Sankaranarayanan[2], Phillip Isola[3]
[1]Department of Computer Science and Maths, Beloit College
[2, 3]Department of Electrical Engineering & Computer Science, MIT


CLIP-Guiding Generative Algorithm

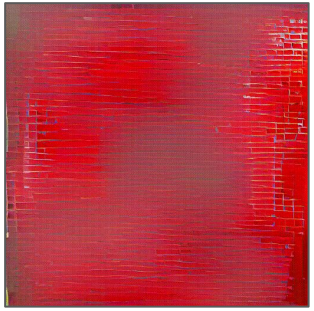
- Recent works in multimodality systems have enabled cross modality generation



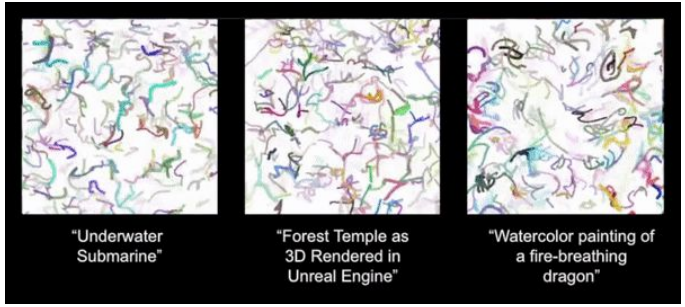
Text2Mesh, Michel et al, 2021

🥓🌆🌊 🏛️❄️🐱

CLIP Guided Diffusion, MidJourney



CLIP + VQGAN, nerdyrodent

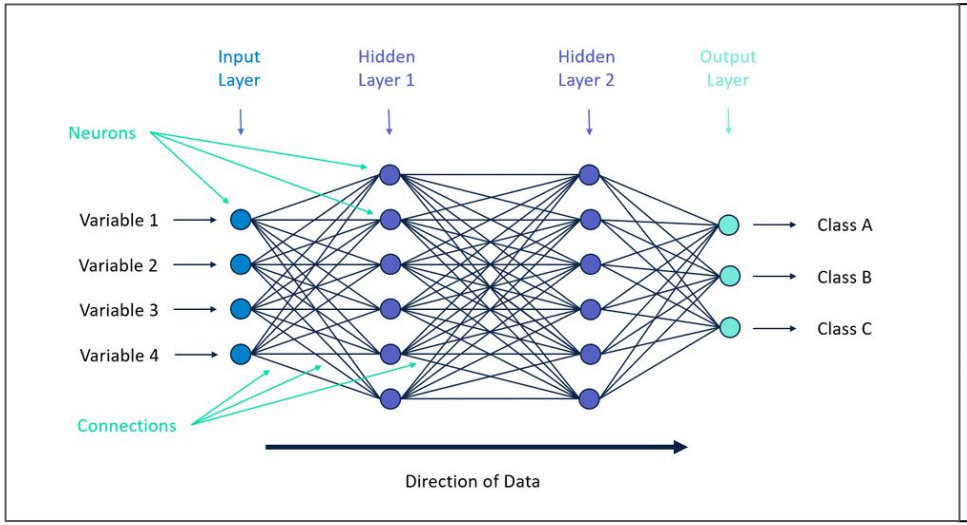


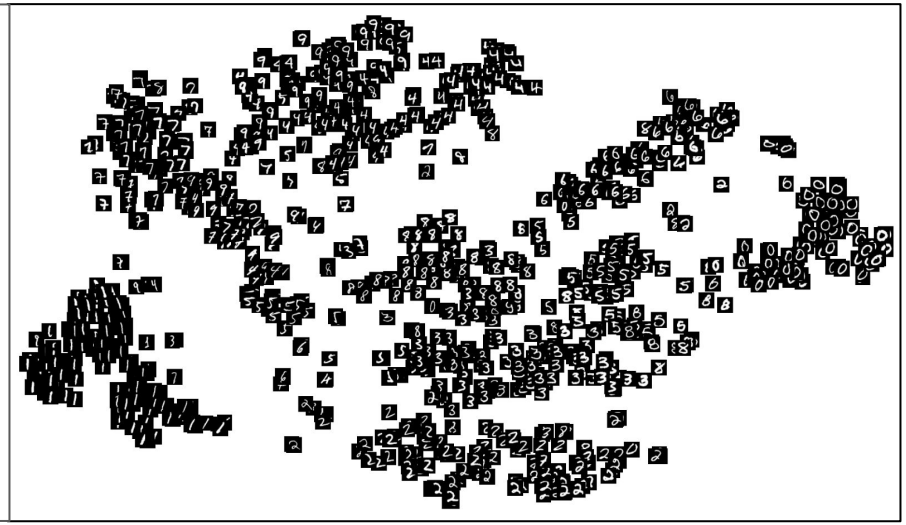
CLIPDraw, Frans et al

Pixray, dribnet

Neural Network and Representation

Representation : The second to last feature vector of neural network



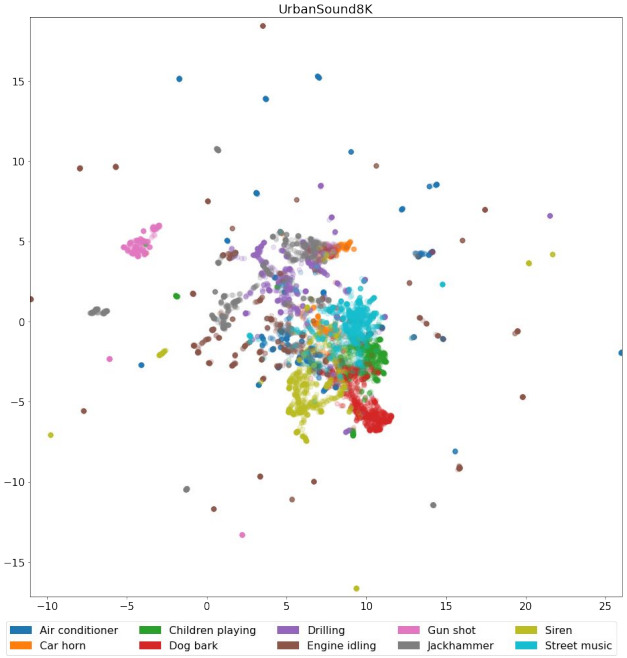


- Representation of the same object is clustered together
- The purpose of contrastive training is to further encourage representations of the same class to be near each other and push other classes' representations far away

Latent Space Visualization, HackerNoon
It's a No Brainer: An Introduction to Neural Networks, Alteryx.com

WAV2CLIP: Extending CLIP to Music

- This is a distilled model from CLIP
- It embed audio to the joint representation space of CLIP
- Could perform multiple downstream tasks

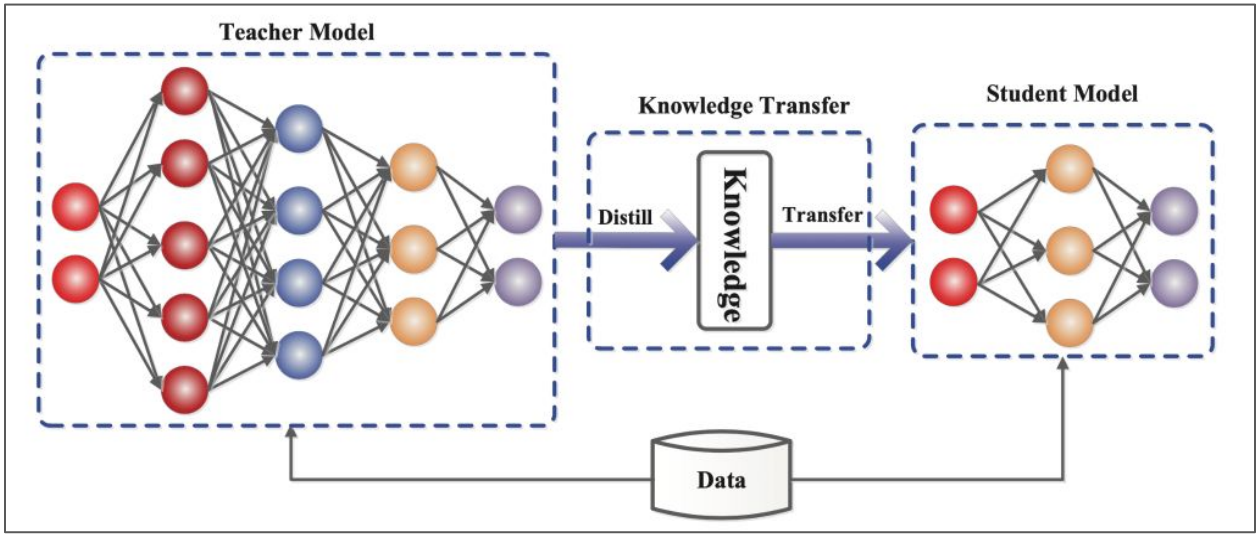


Lyrebird-wav2clip, GitHub

Knowledge Distillation, neptune.ai

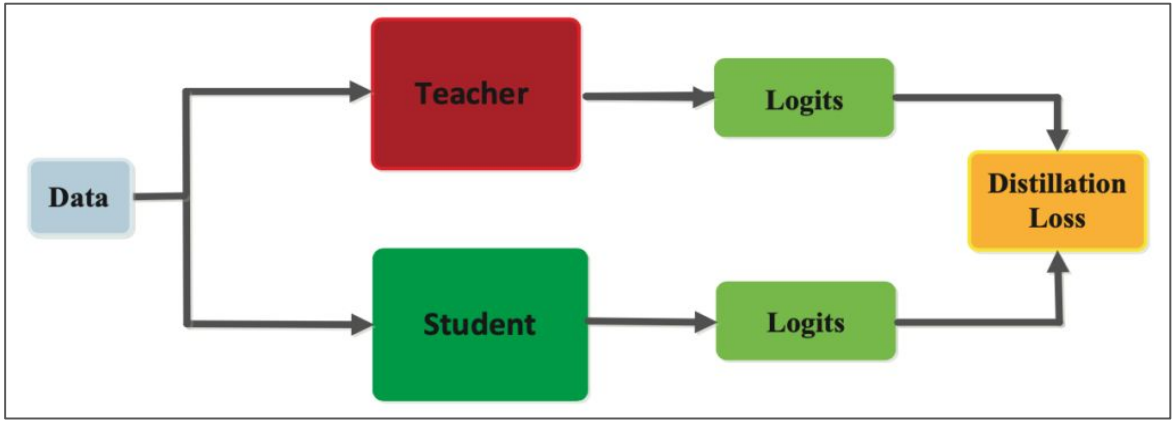
Knowledge Distillation

- Used to transfer knowledge from a bigger model to a smaller model
- Teach model to learn compressed representation without the loss of validity
- Tackle limited memory and computational capacity issue when deploying big deep learning model

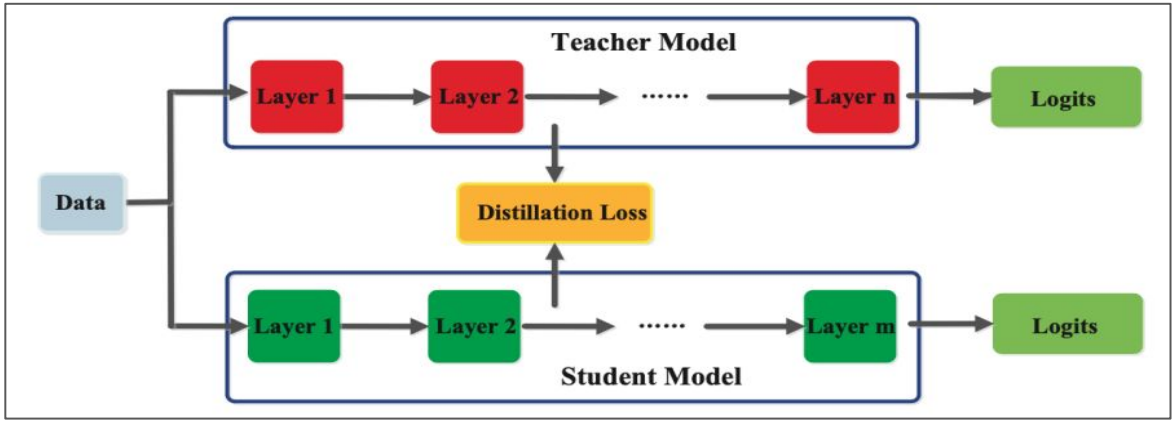


Distillation Type

Knowledge Distillation, neptune.ai



Response based



Feature based

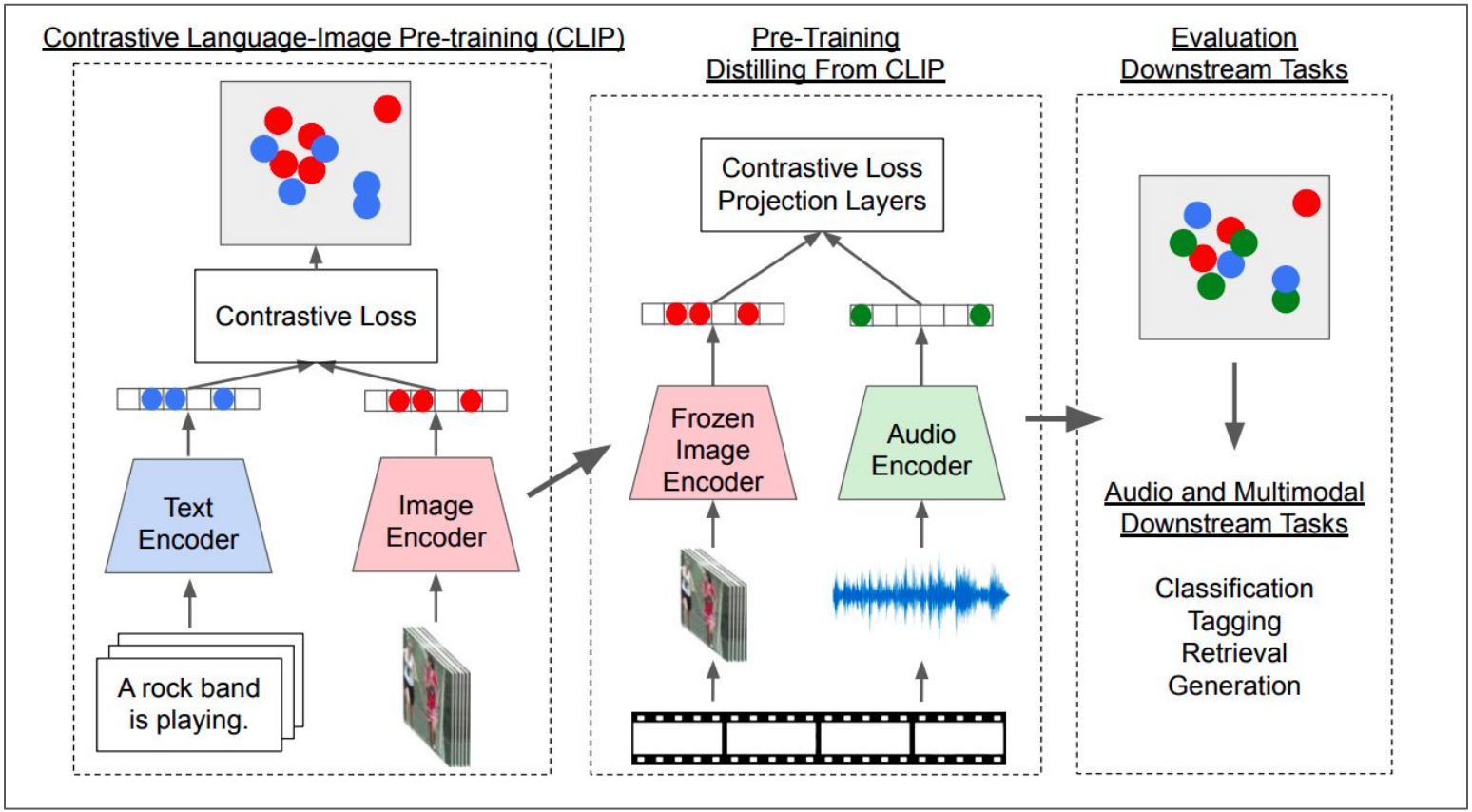
VGGSound Dataset

- For multimodal learning, we will need a dataset includes different modalities
- VGGSound is audio-visual Youtube video dataset. It includes:
  - ~200k 10-second clips
  - 309 class
- For distillation, they sample:
  - 5-second videos that has 150 frames each
  - Get CLIP embeddings for each frames

VGGSound, robots.ox.ac.uk

WAV2CLIP: Learning Robust Audio Representations From CLIP, Wu et al, 2022

WAV2CLIP Distillation Process



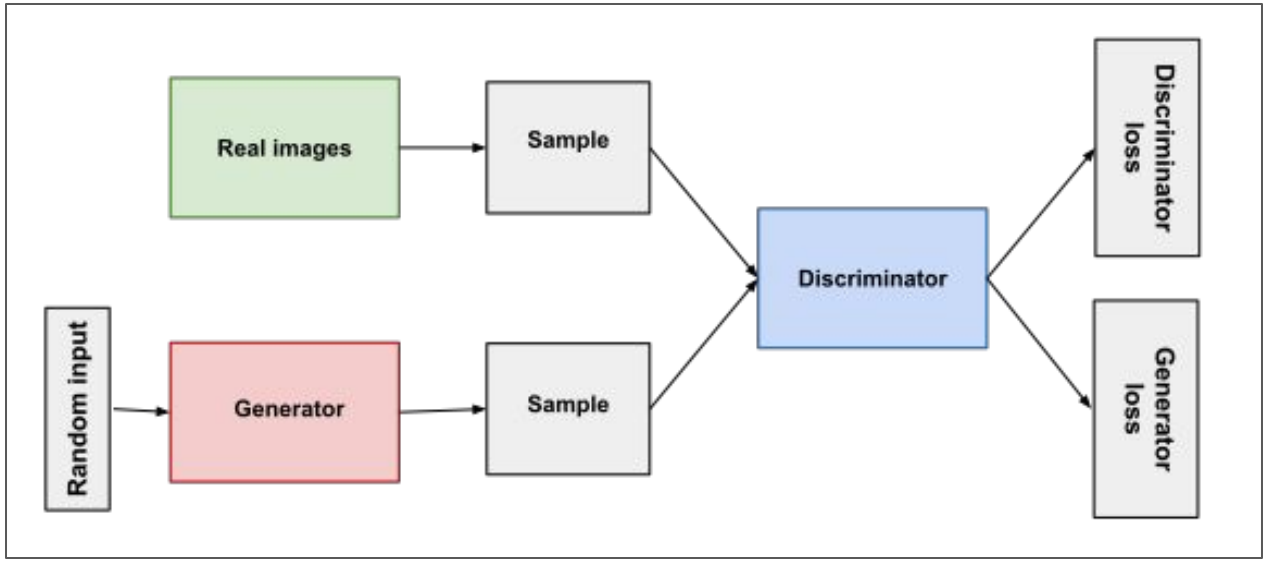
CXLoss = L(f(Image), Audio) + L(Image, g(Audio))

WAV2CLIP: Learning Robust Audio Representations From CLIP, Wu et al, 2022

Sound to Image Generation

Question: Could we make use of the sound representation to generate more meaningful images?

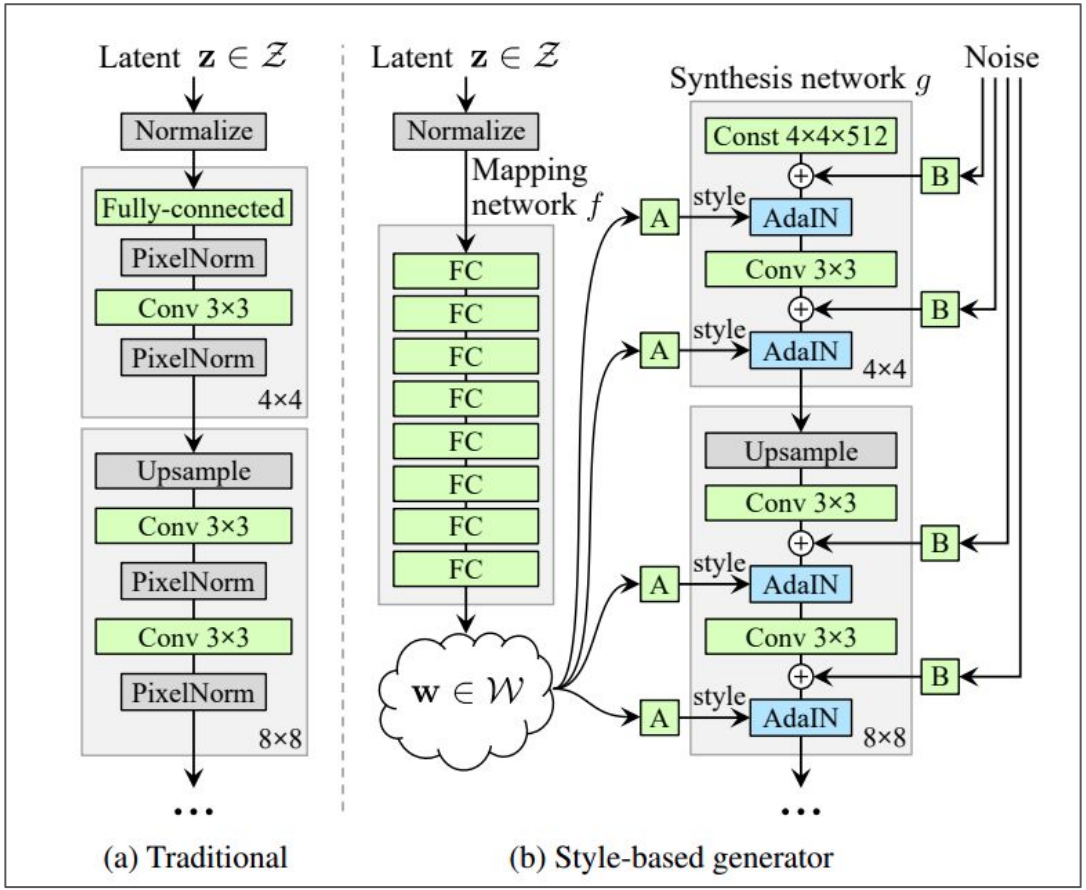
GAN Structure



Overview of GAN Structure, Google Developers

A Style-Based Generator Architecture for Generative Adversarial Networks, Keras et al, 2019

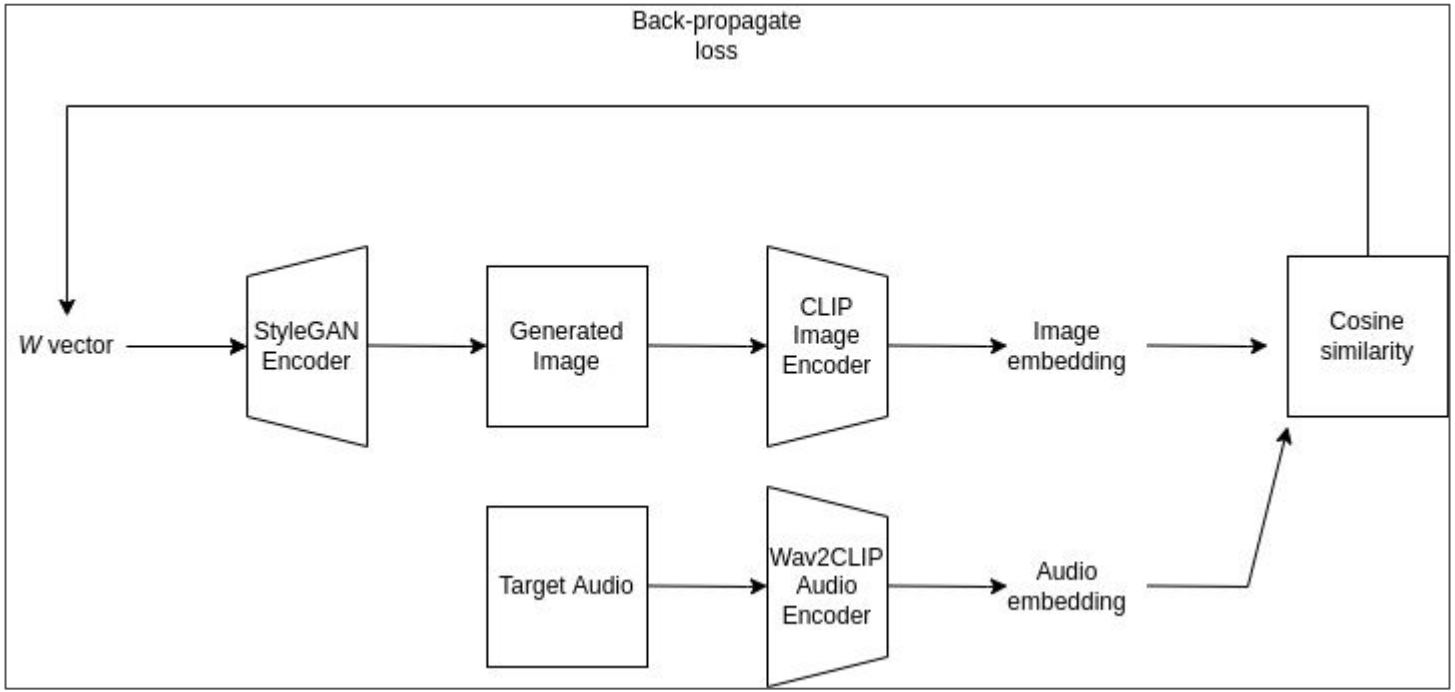
StyleGAN Architecture



StyleGAN Result

https://github.com/justinpinkney/awesome-pretrained-stylegan2

StyleWAV Structure



Loss Function:

$$\arg\min_{w \in W+} D_C(G(w), t) + \lambda_{L2}\|w - w_s\| + \lambda_{ID}\mathcal{L}_{ID}(w)$$

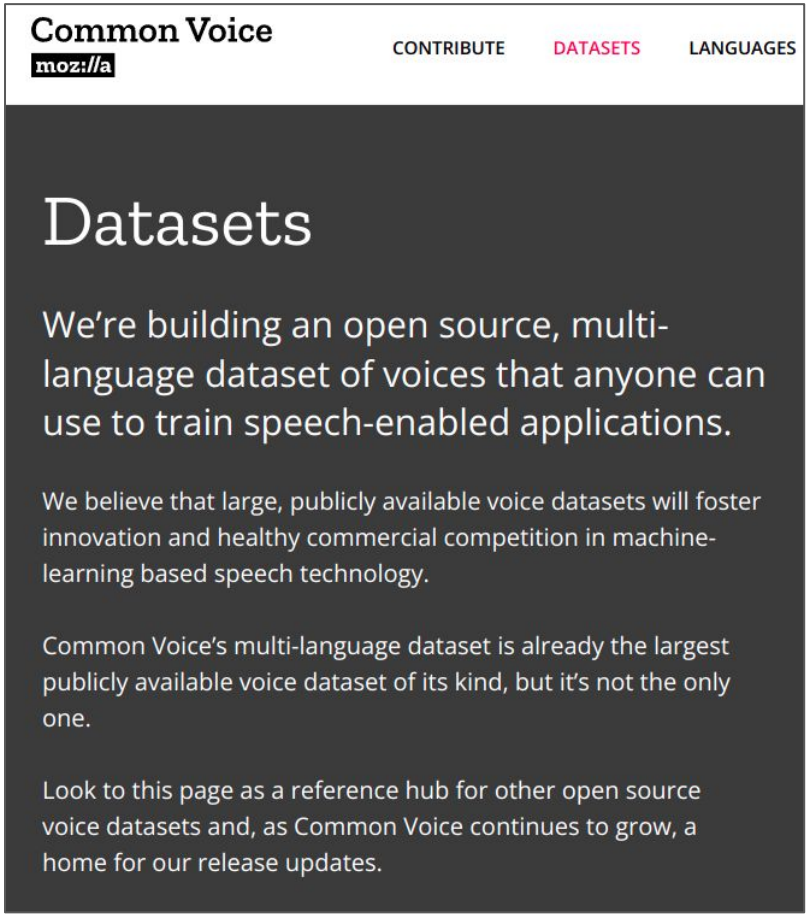
Where G is pretrained StyleGAN, D_C is the cosine distance. T is the audio representation. The two lambda is equal to zero when we do free generation.

$$\mathcal{L}_{ID} = D_C(R(G(w_s)), R(G(W)))$$

R is pretrained ArcFace model for face recognition

Image Generation

- 2 sex: male and female
- Each sex has 100 5-second audios from Mozilla Common Voice Dataset

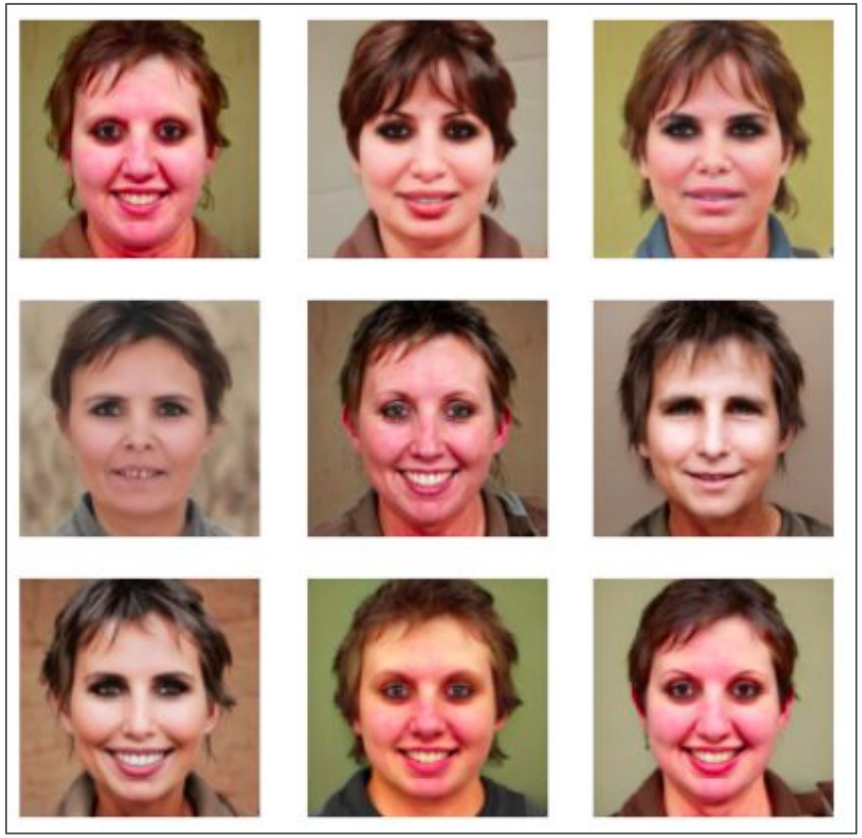


Common Voice, https://commonvoice.mozilla.org/en/datasets

Male Audio

Female Audio

StyleWav Snapshot



T-test

- Perform Welch's t-test on two population of similarity score between two prompts and the same material
- Two mean are significantly different if p-value <= 5%

	Man Photos	Female Photos
this is a photo of a boy	0.26	0.241
this is a photo of a girl	0.252	0.256
have significantly two different means	yes	yes

Thank you for listening

ngoph@beloit.edu