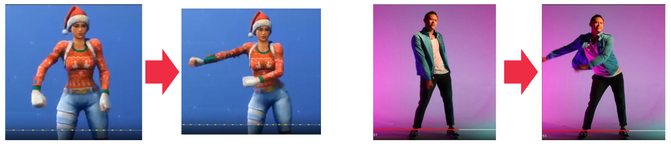
FLOSSLOSS: GAN for Image to Image translation

Annie Chen, Joel Joseph Dominic, Koh Liang Ping

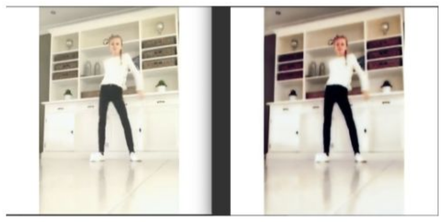
Problem Statement

- Human pose synthesis presents a unique challenge for generation algorithms.
- We used several GAN models to transform standing poses to the floss poses.



Model 1: CycleGAN

- CycleGAN is an unpaired image to image translation model.
- The generator network uses 6 Resnet blocks. We use 64 generator filters in the last convolutional layer and 64 discriminator filters in the first convolutional layer.
- For the discriminator, we use a 70x70 PatchGAN, classifying whether 70 x 70 overlapping image patches are real or not.
- 100 epochs, lr 2e-4, batch size 1:



- The CycleGAN was generally unable to pick up on the change in position, so images generated using this method mostly resembled the original image. However, this makes sense because in the paper, this technique was used mostly for style-transfer related tasks (like changing a photo to different artists' style) and not for making substantial structural changes in images.

Conclusion

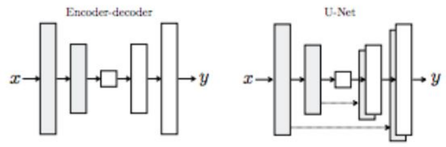
- Last model shows great promise because of the numerous features for facial and pose transfer.
- Some signs of model working in video, although major difficulties in outdated repositories, buggy code and heavy computational cost of training on videos
- Future directions include full implementation of this model with more resources, isolating sub parts of this model to work for images, and using other pre trained models to assist in sharper translation of textures and clothing.

Dataset and Pre-processing

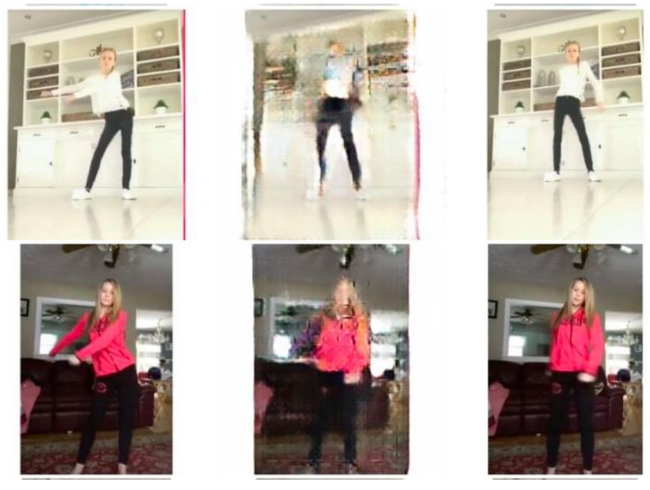
- No prior dataset available.
- Sourced for videos on youtube of people flossing, and then screenshotted them in a standing pose followed by a flossing pose.
- Gathered a total of 135 total images and a total of 55 people, then resized all of the images to be 500 x 500 pixels.
- Augmented our dataset using cropping and flipping.

Model 2: Pix2Pix

- Pix2Pix is a paired image to image translation model, that learns the loss function needed to translate a problem.
- It uses a conditional GAN architecture, whereby rather than generating any mode of data, the generator learns to generate a fake sample with a specific condition.
- It uses a Unet architecture with skip connections between mirrored layers in the encoder and decoder stacks. This is to enable the generator a means to skip the bottleneck for information flow since there is a lot of shared low-level information between the input and output image.



- Tuned hyperparameters: learning rate 0.0004, beta 0.6, batch size 10 lambda 60. Then, we trained for 500 epochs:



Baseline Model

- Vanilla GAN that generates realistic flosses.
- Loss minimized: Binary cross entropy loss

Model 3: Everybody can dance now

- Problems with Pix2Pix: Faces and body parts blurred.
- Hypothesis: Blurring happens because people have different faces, body dimensions, camera positioning. There is no generic mapping that can be learnt for these. Also, whole body parts are not identified and moved together.
- Utilize GAN model that learns from a video to generate video of floss from base image. Has pose identification, pose normalization and facial transfer features.
- Model setup with pose identification/normalization

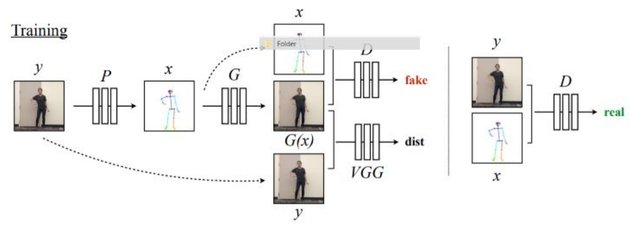


Figure 2. Training architecture for Image to Motion

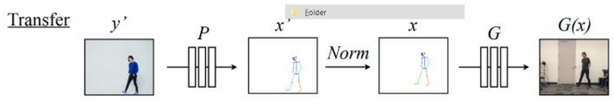


Figure 3. Transfer architecture

- Temporal smoothing loss was used to make video less choppy.
- Facial recognition was also done with a specialized GAN setup.
- Learnt left video to make right image dance. Result in video!