FSGAN: An efficient framework for high fidelity face swapping

[1] Neppalli Bala Krishna Prasad, 11940730, neppallib@iitbhilai.ac.in
[2] Srilekha Kadambala, 11941190, kadambalas@iitbhilai.ac.in

Abstract. Unlike the previous technologies like Deep Fakes, FSGAN is a subject agnostic and can be applied to pairs of faces without requiring training on those faces. We aim to create a new image based on the target image, where target face is replaced by source face without changing the pose and expression of target face. We propose to break large pose changes into small manageable steps and interpolate between the closest available source images corresponding to a target's pose. The Face Swapping GAN (FSGAN) is end-to-end trainable and produces photo realistic, temporally coherent results.

Keywords: U-net · Segmentation · Face reeanctment · pix2pixHD

1 Introduction

Face swapping is the task of replacing a face from source to target image, where it replaces the face appearing in the target image and produces a realistic result. Face reenactment face or puppeteering uses the facial movements and expression deformations of a control face in one video to guide the motions and and deformations of a face which is appearing in another video. This task is attracting significant research attention due to their applications such as in entertainment for visual media production, graphics, pattern recognition, in privacy for photo realistic face de-identification, exchanging faces in images and training data generation. The earliest face swapping methods require manual involvement. After few years an automatic model is proposed. More recently, face to face transferred expressions from source to target face. GAN's will generate fake images with the same distribution as a target domain. Training GANs is somewhat unstable and may produce low-resoluton images. However subsequent methods improved the stability of training process.

2 Problem Definition

We try to replace the target face with face in source image while retaining the target face's pose and expression. In the below figure, the left image is the source image. Right one is

Figure 1: Source face swapped onto target.

the target video. We replace the face in target video with face in source image without changing the pose and expression of target's face. The result is diplayed in the middle.

3 Objective

Let I_s be the source image and I_t be target image with respective faces $F_s \epsilon I_s$ and $F_t \epsilon I_t$. Now we try to create new image based on I_t, replace F_t with F_s while retaining pose and expression. There are mainly three components in FSGAN. The first includes reenactment generator G_r and segmentation CNN G_s. G_r is given the facial landmarks of F_t. Using this it generates a reenacted image $I_r \epsilon F_r$. F_r depicts F_t with same pose and expression. It also computes S_r, the segmentation mask of F_r. G_s generates the face and hair segmentation mask S_t of F_t.

The second component includes filling the missing parts of I_r. The reenacted image I_r may contain missing parts with respect to F_t. So we apply face inpainting network G_c using the segmentation S_t to estimate the missing parts and generate I_c.

The final component includes the blending of face F_c into target image I_t to obtain the final result of face swapping. The blending generator blends the complete reenacted face and the target's face using the segmentation mask.

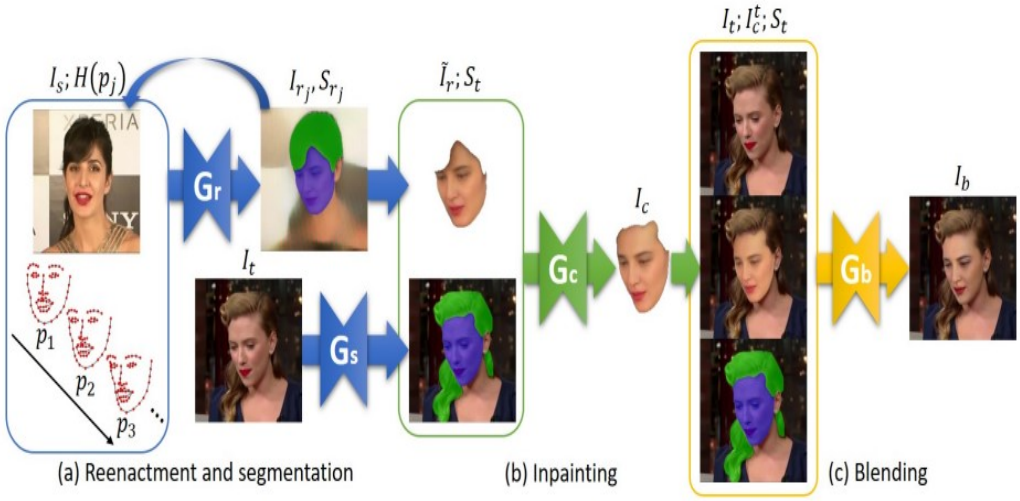

Figure 2: Overview of proposed FSGAN approach

4 Technology used

The face segmentation,G_s, is based on U-net: Convolutional networks for biomedical image segmentation, with bilinear interpolation for up-sampling. All other generators such as G_r, G_c and G_b are based on pix2pixHD. Our global generator uses U-Net architecture instead of simple convolutions and summation instead of concatenations. We use bilinear interpolation for upsampling in both global genrators and enhancers.

Other Python Dependencies used:

- PyTorch

- ffmpeg
- SciPy
- OpenCV
- Pillow
- Torchvision

5 Problems faced

Face swapping in case of complex expressions like screams and surprise. Also, the case where beard or moustache are present on the faces, face swapping gets difficult and unclear. When the source covers their face with their hand or anything in the middle of the video, then again the swapping at that particular frame is unclear.

The training subject of agnostic face reenactment is non-trivial. So there is chance it might fail when applied to unseen face images related by large pose.

6 Data Sets used

In order to train the generator, video sequences of the IJB-C dataset are used. IJB-C contains approximately 11k face videos of which only 5,500 were used for the training purpose that are in high definition. The dlib's face verification is used to group frames according to the subject identity, and limit the number of frames per subject to 100, by choosing frames with the maximal variance in 2D landmarks. In each training iteration, the frames I_s and I_t are choosen from two randomly chosen subjects. The VGG-19 CNNs are trained for the perceptual loss on VGGFace2 dataset for face recognition. The CelebA dataset are used for face attribute classification. And also the LFW parts labels set with approximately 3k images labeled are used for face and hair segmentations. Additional 1k images from the Figaro datasets are used. Finally, FaceForensics++ provides 1000 videos, from which they generated 1000 synthetic videos on random pairs using DeepFakes and Face2Face.

7 Models Used

There are 2 models mainly used in this FSGAN approach for face swapping. U-Net and Pix2pixHD. UNet is a convolutional neural network architecture that expanded with few changes in the CNN architecture. It was invented to deal with biomedical images where the target is not only to classify whether there is an infection or not but also to identify the area of infection. The two main parts of Unet architecture are encoder and decoder. The encoder is used to extract factors from the image whereas decoders is uses transposed convolution to permit localization. Image segmentation is a process where the image is segmented into different segments representing each different class in the image. When an image is given as input to Unet architecture, the segmentation map of that image is given as output. In our FSGAN approach, G_s uses Unet architecture. When the target image(I_t) is given as input for Unet, it generates the face and hair segmentation mask of the target image. So the segmentation map of target image is generated using Unet model in our Face swapping.

The second model is pix2pixHD which is used by generators such as G_r, G_c and G_b. Pix2pixHD used GANs for high resolution image-to-image translation by applying a multi-scale conditional GAN architecture and adding a perceptual loss. It can generate high

resolution image results. In less technical terms, pix2pixHD is a straightforward way to generate high-resolution images with nearly endless options to change small and large details about the images. This is done by drawing on a label map and then translating the drawings using GANs to produce HD image outputs. In case of G_r, using the facial landmarks of target image, a reenacted image is generated. Missing parts of reenacted image is filled and G_c generates the full face image.G_b blends the image generated by G_c and target image to generate the final face swapped image. So to have a high resolution to all the generated images and also to have the skin colour of the target's face preserved, pix2pixHD is used un this FSGAN approach of face swapping.

8 Result and Performance

Figure 3 shows the face swapping results of different approaches. These examples are taken from FaceForensic++ videos which is included in our dataset.

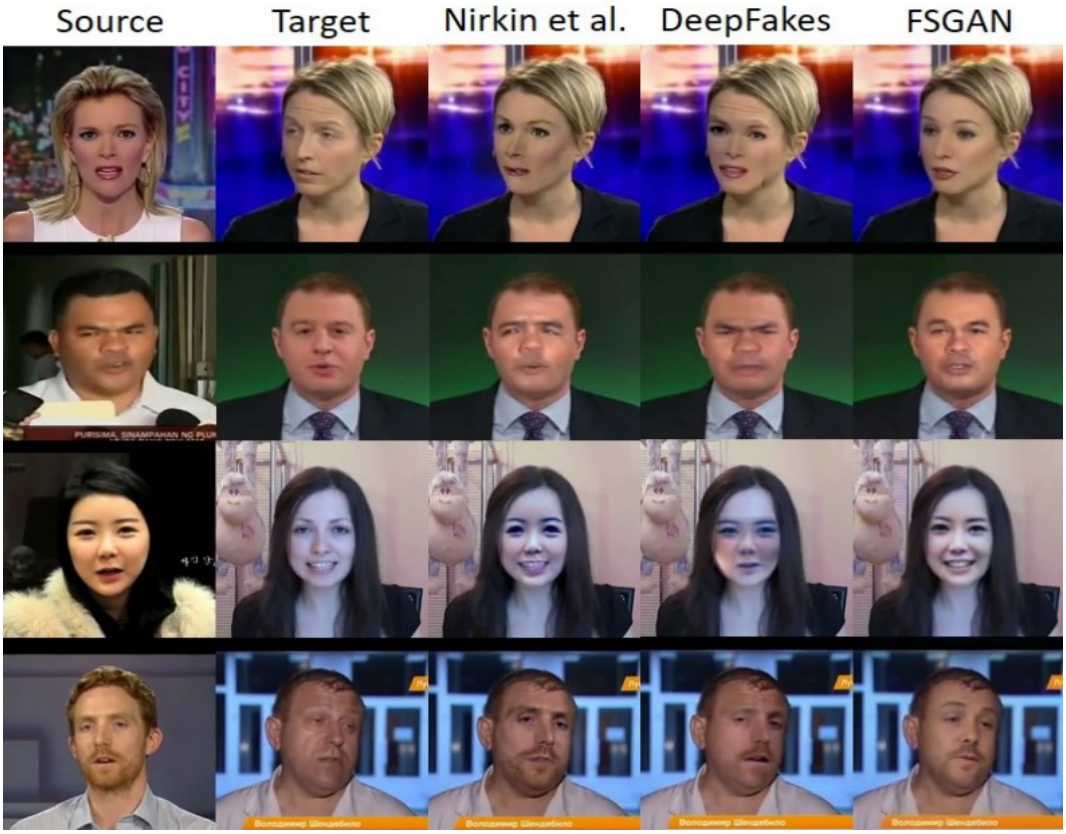

Figure 3: Results for source photo swapped onto target provided for Nirkin et al., DeepFakes and our method on images of faces of subjects it was not trained on.

The examples are choosen in such a way that they have different poses and expressions, face shapes. For each frame in the target video we select the source frame with the most similar pose. When we compare our reenactment result with that of the Face2Face result, this is shown in figure 4. Given a source face and target face, the goal is to transfer the expression of source face to target face. The modification is done in such a way that the corresponding 2D landmarks of target face is changed by swapping in the mouth points of

the 2D landmarks of source face. If we observe from figure 3 and figure 4, not only in face swapping, but also in reenactement the results obtained using FSGAN are much better than the previous works like Face2face and DeepFakes.

Figure 4: Comparison to Face2Face on FaceForensics++

9 Conclusion and Further Work

We can validate how well the methods preserve the source subject identity, while retaining the same pose and expression of the target subject. The FSGAN model has outperformed many other models without even training subject-specific images. GANs are successful in generating fake faces from realistic images. The FSGAN method eliminates subject-specific, data collection and also training on the model. This makes the face swapping and reenactment easy to learn. Overall, Face Swapping GANs are really an impressive method of GANs.

We have done video to video swapping as of now. Image to image swapping has done but not attached with the code submitted as we haven't finished the image to image swap yet. Also, we will be doing image to video face swap and video to image face swap.

10 References

1) https://analyticsindiamag.com/gans-can-swap-faces-now-with-fsgan-a-new-deep-learning-approach-to-face-swapping/
2) https://golden.com/wiki/Pix2pixHD
3) https://medium.com/@ODSC/fsgan-subject-agnostic-face-swapping-and-reenactment-2f033b0ea83c
4) https://analyticsindiamag.com/my-experiment-with-unet-building-an-image-segmentation-model/
5) https://arxiv.org/pdf/1908.05932.pdf