AML Paper Review - **Subhani Shaik (MT18117)**

FSGAN: Subject Agnostic Face Swapping and Reenactment

Authors: Yuval Nirkin, Yosi Keller, Tal Hassner

I. SUMMARY

Face swapping is a task of transferring a face from source to target image. Face Reenactment means involving facial movements and expressions based on appearance of source image to target image using transformations. FSGAN (Face Swapping GAN) is an end-to-end trainable architecture for swapping and for re-enactment of the face from source to target image. There are many other approaches which worked on the same problem like 3D face representations for face swapping, DeepFake architecture, latent feature based architectures, disentanglement and GANs which generates face with respect to subject. The main advantage of this architecture is- it is subject agnostic i.e., we can transfer to face along with re-enactment without training that subject.

To get better results authors have used Perceptual Loss L_{perc}, Pixelwise loss (L1 loss) L_{pixel}, Reconstruction loss L_{rec} which is combination of L_{perc} and L_{pixel}, and Adversarial loss L_{adv} which is GAN loss.

The FSGAN has three main components- Face Reenactment and Segmentation, Face Inpainting and Face Blending.

A. Face Reenactment and Segmentation

In this, G_r (re-enactment generator) which takes an image *I* and a heatmap *H(p)* which has facial landmarks. G_r generates the re-enacted image I_r and segmented image of hair and face S_r with the help of G_S (Segmentation CNN- UNet), which further used this intermediate generated image to generate enhanced image with face which might also have missing pixels. G_S is used to analyse and generate the segmented hair and face image S_t of target.

Here for training authors have used Stepwise consistency loss for G_r and cross-entropy loss for G_S. This network was trained G_r and G_S alternatively. To get better interpolation of face for given subject images $\{I_{s1..sn}\}$, Euler angle $\{e_{1..n}\}$and Faces$\{F_{s1..sn}\}$we find the closest triangular plane position of target image and calculate barycentric coordinates and generates S_r.

By end of this component we will have I_r, S_r, S_t which are further send to next component.

B. Face Inpainting

Generally re-enacting the occluded faces is difficult and more chance to generate artifacts. To deal with that problem authors have proposed this Inpainting generator G_c. It takes I_r, S_r and I_t and render the better context with respect to target image by randomly removing elliptical shape parts like hair etc. It used generator loss

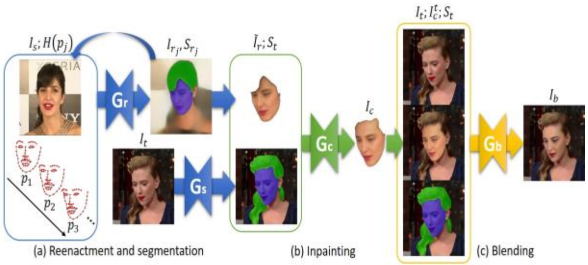
$$L(Gc) = \lambda recLrec(Ic; It) + \lambda advLadv.$$

We will get F_c image as output which is further send to next component.

C. Face Blending

This component take F_c and blends with respect to F_t. To take skin-tone and lighting conditions into account authors have used Poisson blending loss. For generating better re-enacted face authors have combined this poisson blending loss with perceptual Loss L_{perc}.

$$L(Gb) = \lambda recLrec(Gb(It; Irt; St); P(It; Irt; St))+\lambda advLadv.$$



II. LIMITATIONS OF FSGAN

The FSGAN works better than earlier approaches like DeepFake etc. and have advantage like subject agnostic network, but it has some limitations:

- For large images with occlusions, at the time of inpainting it may generate artifacts which can result in blurriness and degrade the quality of re-enacted image.
- Because of iterative process it may generate degraded texture of the image.
- It wraps the texture directly from the target images i.e., more dependent on attributes of training data like resolution etc.
- The authors have used facial landmarks for re-enactment with can be sparse, because of that architecture may not capture facial expressions correctly.

III. SUGGESTIONS FOR THE PROPOSED TECHNIQUE

- For segmentation authors have used UNet with bilinear interpolation for upsampling. We can use segmentation architectures like FastFCN, Gated-SCNN which give better results. So that it will improve quality of re-enactment.
- We can use occlusion removal architecture like 3DMM which will perform better images by wrapping texture from the images itself.
- For better facial land marking we can use RNN with attention architecture in re-enactment component. So that it can generate better image I_r depicting that features of I_t.

IV. REFERENCES

[1] FSGAN: Subject Agnostic Face Swapping and Reenactment. Yuval Nirkin et.al.

[2] FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation. Huikai Wu et.al.

[3] Gated-SCNN: Gated Shape CNNs for Semantic Segmentation. Towaki Takikawa et.al.

[4] Face De-occlusion using 3D Morphable Model and Generative Adversarial Network. Xiaowei Yuan et.al.