Milestone 1 Report

Project Proposal

Presented to

Mashhour Solh

San José State University

Computer Science Department

CS256, Topics in Artificial Intelligence, Spring 2022

By

Ajinkya R, Branden Lopez, Indranil Patil, Rushikesh P, Warada K

3/2022

Table of Contents

I. PROBLEM/BACKGROUND

Computer Vision (CV) is an interdisciplinary field aiming to solve computers' understanding of digital images and videos. Specific CV tasks include Motion Transfer, Image Generation, and Style Transfer.

Motion transfer tracks an object's motion over time and applies the same actions to a candidate image. Image generation aims to create new realistic images. Finally, Style transfer takes two images, a content image, and a style reference image, then blends them so that the content image has the stylization of the other.

The above tasks often overlap and have many uses in the industry. The current state-of-the-art methods in motion transfer only work on a single target and our contribution aims for the animation of two targets.

II. DATASETS

There are two free and publicly available datasets-

1. VoxCeleb2 [1]: Audiovisual dataset of millions of videos of celebrities extracted from YouTube videos. Videos contain a variety of poses, backgrounds, and lighting conditions.

2. TalkingHead-1KH [2]: Contains thousands of hours of video recordings of celebrities and people from varied sources with open licenses. Videos are generally of higher quality and resolution than VoxCeleb2.

III. METRICS

1. PSNR: Approximates to human perception of reconstruction quality by using peak signal-to-noise ratio.

2. SSIM/MS-SSIM: SSIM is more robust than PSNR. It measures structural similarities between the patches. MS-SSIM is a multi-scale variation of SSIM.

3. Fréchet Inception Distance (FID) [3]: Measures how close is the distance between the distributions of synthesized and real images.

4. Average keypoint distance (AKD): Using a facial landmark detector to detect landmarks of real and synthesized images and then compute the average distance between the corresponding landmarks in these two images. Dlib-ML library provides facial landmark detectors for the detection [4].

IV. CLASSICAL APPROACH

One of the applications for domain adaptation or image generation can be to split two component images and merge them. An example of this can be seen in Figure (1) where the left half of an apple is merged with the right half of an orange. [5] has described a multiresolution spline technique that first decomposes the component images into a set of band-pass filtered images and then the images in each frequency are assembled into the corresponding band-pass mosaic.

Lastly, these band-pass mosaic images are then added to get the desired image. The working of this technique is explained in Figure (2). In this implementation, for splining and filtering, pyramid algorithms have been used. It is known to be an efficient filter as it has only seven arithmetic operations including additions and multiplications per pixel. The pyramid algorithms offer an easy and efficient way in which filtering and splining can be performed.

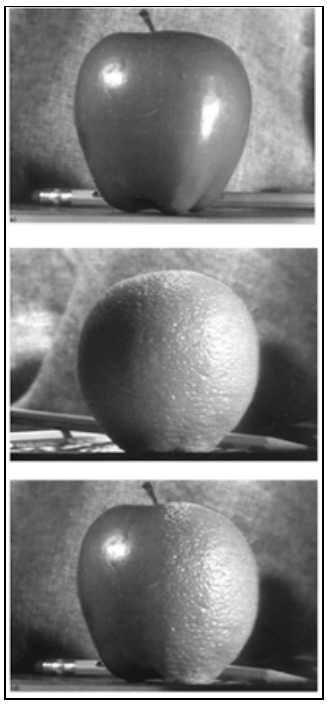

Figure 1. Left half of an apple combined with the right half of an orange

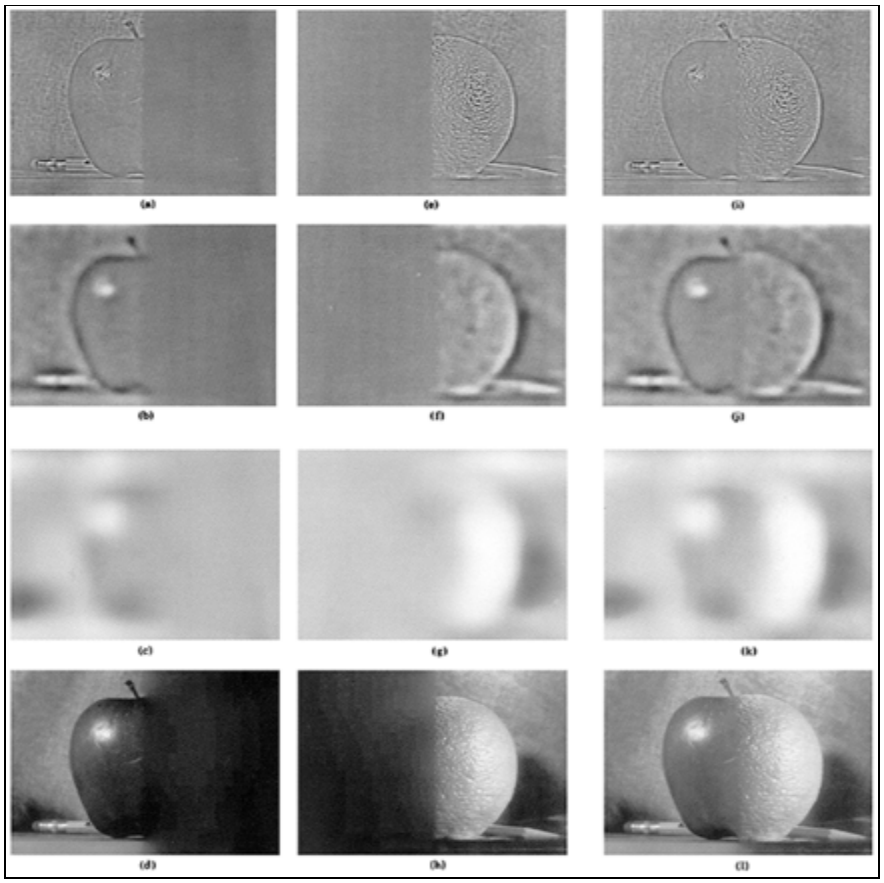

Figure 2. Contributions of the various band-pass filtered images to the apple-orange mosaic. Figure 10a-10c shows low, medium, and high frequency components of apple. Figure 10d is obtained by summing the figures 10a-10c. Same frequency components are obtained for orange in the Figures 10e-10h. Images in figure 10i-10l are obtained by summing the frequency components of half apple and half orange.

V. STATE OF THE ART

A. CNN

The problem of style transfer can also be referred to as texture transfer. [6] introduced *A Neural Algorithm of Artistic Style* that uses the feature space derived by the VGG network. The proposed algorithm enables the generation of new images which combines the content of a photograph with the style of artworks as depicted in figure (3). The functioning of the algorithm is as illustrated in figure (4). However, a good tradeoff between the style and content that needs to be obtained is difficult to synthesize

Figure 3. Images combining the content of A with artworks B, C, and D

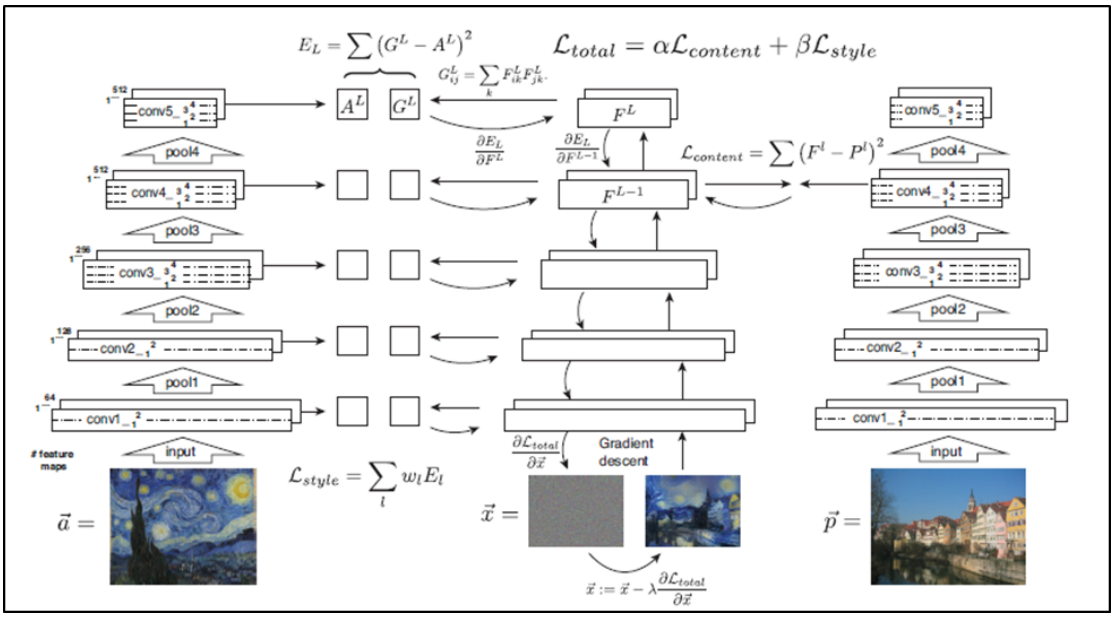

Figure 4. Style Transfer Algorithm

B. Conditional GANs and CycleGANs

The issue with the above approach was translation of output images were mostly a combination of two images rather than actual style transfer. The issue was solved by using Conditional Generative Adversarial Networks. Image to image translation can be considered as a special case of Conditional GANs [7]. However, in this case, instead of using a single label for conditioning on the image, we are using a complete image as an input label [8].

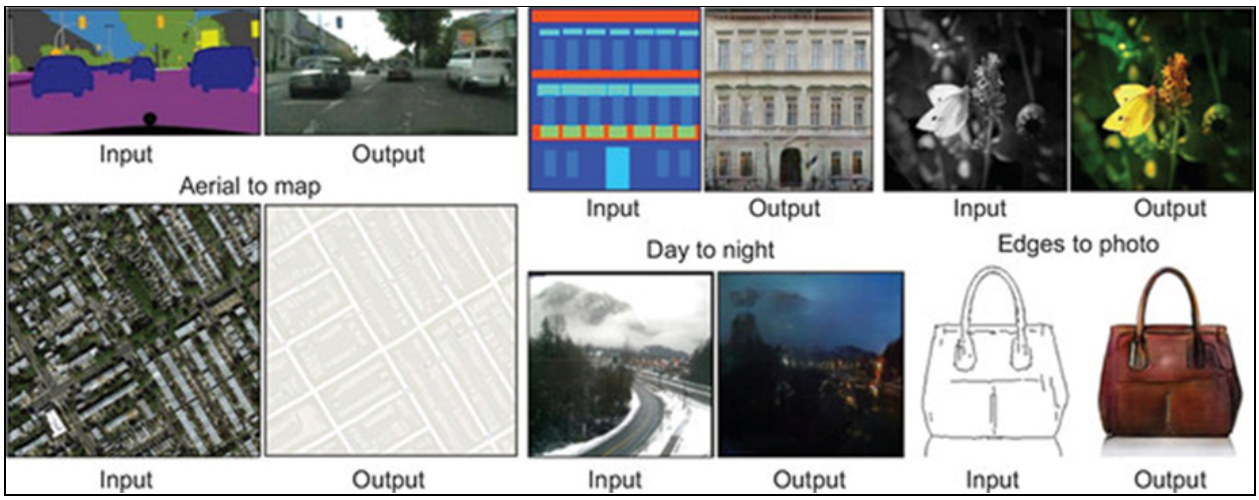

Figure 5. (Source: "Image-to-Image Translation with Conditional Adversarial Networks," by Phillip Isola [8])

The main drawback of this approach is that it requires "perfect" pairs of images from both domains to train models. To tackle this, CycleGAN introduced an additional loss function called cycle-consistency loss[9]. We simply complete the cycle: we translate from one domain to another and then back again. The reconstructed image must be the same as the original image. The difference between original image and reconstructed image is used to compute the cycle consistency loss.

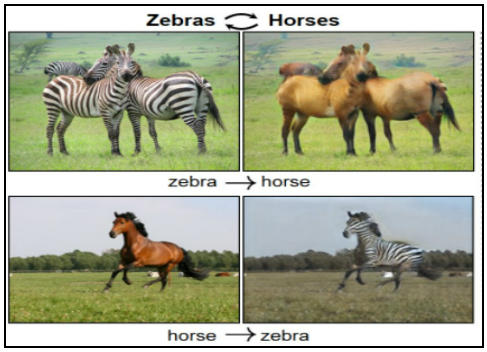

Figure 6. Cycle-GAN example. We go from zebra picture (domain A) to horse one (domain B) and then back again to zebra (domain A). (Source: Jun-Yan Zhu et al., 2017 [10])

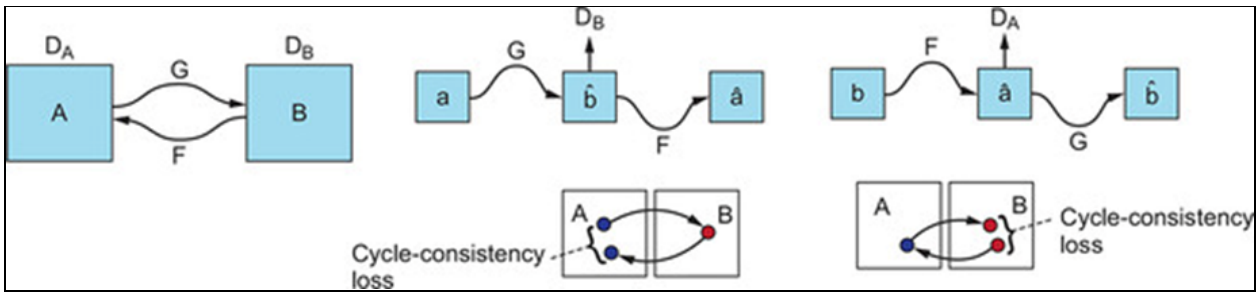

Figure 7. (Source: Jun-Yan Zhu et al., 2017 [10])

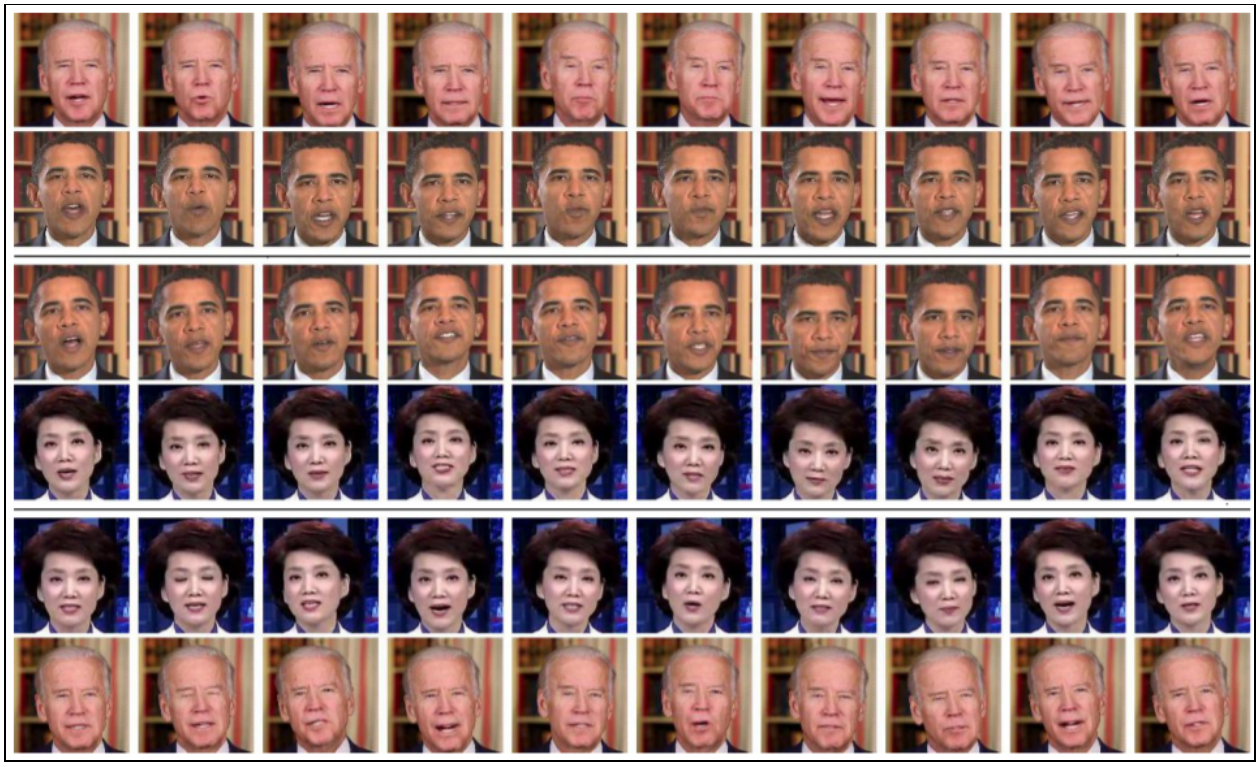

Figure 8. (Source: Face Transfer with Generative Adversarial Network [11])

C. VAE and Coupled GANs

Alternative to the approach mentioned above, [12] proposed an algorithm of unsupervised image-to-image translation that aims at learning the joint distribution of images in separate domains by using images from marginal distributions in individual domains. Since nothing can be inferred from the infinite possibilities of joint distributions, an assumption for two images sharing a latent space such that both these images can be recovered from the latent space and vice versa is being made. The framework is based on virtual encoders (VAEs) and

coupled generative adversarial networks (GANs), where learning of translation takes place in both directions in one shot. The proposed framework was applied on the CelebA dataset with images having facial attributes (eyeglasses, smile, blond hair, etc.) under 1st domain and without facial attributes in 2nd domain. The attributes were then translated from domain 1 to the images of domain 2 as visualized in figure (9), without having any corresponding images in both the domains in the training dataset. However, this framework could be unstable due to the saddle point searching problem.

Figure 9. Attribute-based face translation results

D. Style GAN

Traditional GANs offer very limited control over the newly generated images, Karras et al. [13] introduce StyleGAN, a new generator that learns to separate different aspects of the image. The new architecture perceives an image as a collection of styles where each style controls the effects at a particular scale and adjusting the image style at every layer allows control of image features at multiple scales; allowing the style modification to only affect a certain aspect of the target image. This implementation has classified the styles in three different sections namely – coarse style – which includes pose, hair, face shape, Middle styles – facial features, eyes, and Fine Styles – color scheme. The Gaussian noise introduced at each layer of the network with the style changes improves the unsupervised separation of high-level attributes and interpolation properties. However, the paper does not explicitly mention adopting this architecture for modifying non-synthetic images.

E. JoJoGAN

JoJoGAN [14] is a One-Shot image stylization model built on StyleGan. While art styles are often few, JoJo GAN utilizes GAN inversion to produce a paired data set from a single styling example. Finetuning StyleGAN on a new dataset [15] and performing layer swapping allows the StyleGAN to learn image to image translation with a relatively small dataset, making One-shot stylization possible. While image stylization is controlled via StyleGANs

mixing features, JoJoGAN eye gaze direction is from the reference image rather than the input.

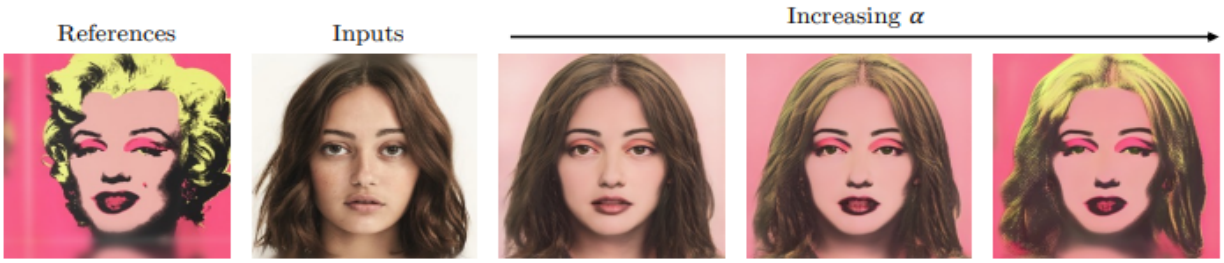

Figure 10: Image Stylization with control over the impact of the reference image.

F. Image Animation -First Order Motion Model

In [16] a source image S is animated based on the motion of a similar object from a video D. By describing the motion as a set of keypoint displacements and local affine transformations, a GAN combines the appearance of S and the motion of D; following a self-supervised strategy inspired by Monkey-Net.

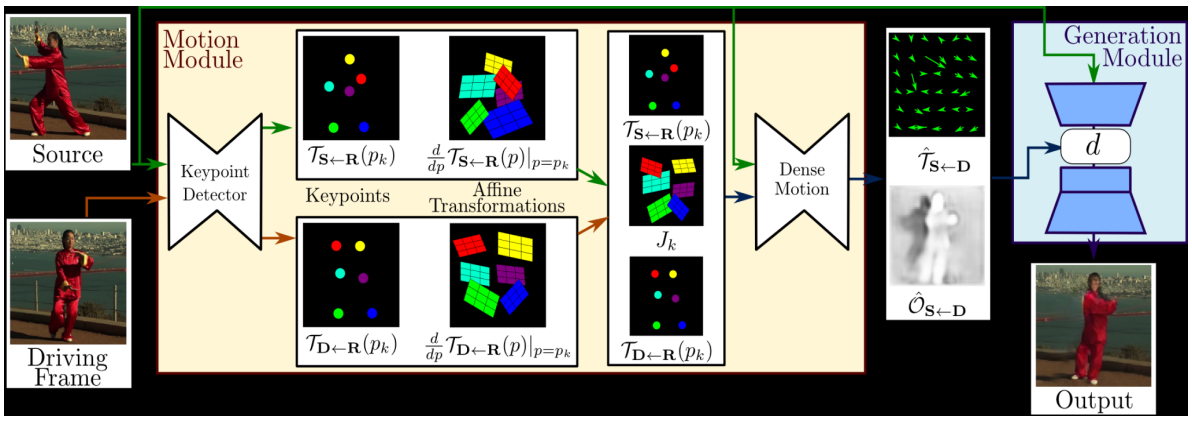

Figure 11: Overview of First Order Motion Model

VI. PROJECT APPROACH

The state-of-art approaches discussed till now use 2-D based image translation. These approaches demonstrate excellent result qualities but can only synthesize fixed viewpoint videos which produce less immersive experiences. 3D model-based talking-head synthesis methods derive deformation from a 3-D source model and apply it to the target model. The main

drawback of these approaches is they do not work well with coarse-grained features like hair, teeth, and accessories.

In [17] Wang et al. propose a new model to generate a video of a talking head derived from a source image (appearance) and a driving video. In this approach, appearance, canonical keypoints, head pose, and expression features are derived from source images using independently trained deep learning models. Similarly, head pose and expression features are derived from driving video. These two features are combined with appearance and canonical keypoints features of the source using a novel 3D keypoint representation proposed by the paper. A generator network is then used to generate realistic images from the combined features.

The paper proposes a technique where fine-grained features derived from the source and target image allow for a much finer level of control over generated images. As Independent features like head pose can be easily manipulated by the user, this approach can also be used for face fractalization. As the size of extracted features from the driving image is much lesser than the image, it allows for more efficient compression with video quality on par with commercial standards while using only one-tenth of the bandwidth.

To contribute to the CV field, we will utilize the state-of-the-art approach of Face2Vid and animate multiple faces. This model is chosen as its one-shot property will adjust to the small datasets provided. Our contributions will utilize a pre- and post-processing of images that will determine where the faces are, separate the entities, process the individual images, and concatenate the animations.

References

[1] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. “VoxCeleb2: deep speaker recognition”. In INTERSPEECH, 2018.

[2] Wang, T.-C., Mallya, A., & Liu, M.-Y. (2021). “One-Shot Free-View neural talking-head synthesis for video conferencing”. CVPR.

[3] Martin Heusel, et al, “GANs trained by a two time-scale update rule converge to a local nash equilibrium”. In NeurIPS, 2017 References

[4] Davis E. King. “Dlib-ml: A machine learning toolkit”. JMLR, 2009

[5] P. Burt and E. Adelson, "A multiresolution spline with application to image mosaics," ACM Transactions on Graphics, vol. 2, (4), pp. 217-236, 1983. . DOI: 10.1145/245.247.

[6] L. A. Gatys, A. S. Ecker and M. Bethge, "Image style transfer using convolutional neural networks," in 2016. doi: 10.1109/CVPR.2016.265.

[7] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” arXiv preprint arXiv:1411.1784, 2014.

[8] P. Isola, et al, “Image-to-image translation with conditional adversarial networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.

[9] Langr, J., & Bok, V. (2019). “GANs in Action: deep learning with generative adversarial networks”. Manning. https://books.google.com/books?id=HojvugEACAAJ

[10] Zhu, J.-Y., et al, “Unpaired image-to-image translation using cycle-consistent adversarial networks”. Computer Vision (ICCV), 2017 IEEE International Conference On.

[11] R. Xu et al, "face transfer with generative adversarial network," 2017.

[12] M. Liu, T. Breuel and J. Kautz, "Unsupervised image-to-image translation networks," 2017.

[13] T. Karras, S. Laine, and T. Aila, “A Style-Based generator architecture for generative adversarial networks,” presented at the - 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4396–4405, doi: 10.1109/CVPR.2019.00453.

[14] Chong. Min, and Forsyth. D.A, “JoJoGAN: one shot face stylization”, 2021 doi: arXiv:2112.11641

[15] Pinkney, J.N., Adler, D.: “Resolution dependent gan interpolation for controllable image synthesis between domains”. arXiv preprint arXiv:2010.05334 (2020)

[16] Siarohin, Aliaksandr, et al. "First order motion model for image animation." Advances in Neural Information Processing Systems 32, 2019

[17] T.-C. Wang, A. Mallya, and M.-Y. Liu, “One-Shot Free-View neural talking-head synthesis for video conferencing,” presented at the - 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10034–10044, doi: 10.1109/CVPR46437.2021.00991.