Real Time Facial Expression Transformation

Ravikant Sharma
Dept of Information Technology V.C.E.T.
Vasai, India

Krithika Suvarna
Dept of Information Technology V.C.E.T.
Vasai, India

Buddhghosh Shirsat
Dept of Information Technology V.C.E.T.
Vasai, India

Prof. Maryam Jawadwala
Asst. prof. Dept of Information Technology
V.C.E.T. Vasai, India

***Abstract*---Facial Expression capture is the process of electronically converting the movements of a person's face into a digital database using cameras or laser scanners. The method is to transfer facial expressionsfrom an actor in the source video to an actor in a target video in real-time. Thus, enabling the improvised control of the facial expression of the target actor. The originality of our approach lies in the transfer of photorealistic re-rendering of facial deformation and detail into target video in a way that the newly synthesized expressions are made as indistinguishable from the real video as possible. To achieve this a trained dataset of images will be created from the target video, then using any webcam or any camera as input feed source's facial traits will be captured and the target face will be manipulated according to that. For each frame, the environment will be considered and ensured that only facial expressions and motions change and everything else remains intact otherwise the viewer might feel that something is out of place.**

***Keywords*---Reenactment, expression, transfer, real-time, source video, target video.**

I. INTRODUCTION

In recent years, real-time markerless facial performance capture based on community sensors has been demonstrated. Impressive results have been achieved, both based on RGB
[1] as well as RGB-D data. These techniques have become increasingly popular for the animation of virtual CG avatars in video games and movies. It is now feasible to run these face capture and tracking algorithms from home, which is the foundation for many VR and AR applications, such as teleconferencing.

Facial Motion Capture is related to body motion capture but is more challenging due to the higher resolution requirements to detect and track subtle expressions possible from small movementsofthe eyesand lips. These movements are often less than a few millimeters, requiring even greater resolution and fidelity and different filtering techniques than usually used in full-body capture. The additional constraints of the face also allowmoreopportunities for using models and rules.

Facial expression capture is similar to Facial Motion Capture. It is a process of using visual or mechanical means to manipulate computer-generated characters with input from human faces or to recognize emotions from a user. And once the facial expression is transferred it can be used in various fields like the gaming world and movies and man more. Facial expression transfer is a difficult task as the background is to be kept stable so that it doesn't look weird to the viewer. And the database needs to be maintained properly so that no problem is faced while transferring the expression from the source to the target actor. A major challenge is the convincing re-rendering of the synthesized target face into the corresponding video stream. This requires careful consideration of the lighting and the shading design, which both must correspond to the real-world environment.

Generative adversarial networks (GANs) for example, were shown to successfully generate realistic images of fake faces. Conditional GANs (cGANs) were used to transform [5] an image depicting real data from one domain to another and inspired multiple face reenactment schemes. These methods decompose the identity component of the face from the remaining traits and encode identity as the manifestation of latent feature vectors resulting in significant information loss and limiting the quality of the synthesized images.

II. PROBLEM STATEMENT

Real time facial expression transformation aims to transfer the expression of the source face on to the target face using various techniques and algorithms. A platform wherein we will be doing real-time source-to-target reenactment approach for complete human portrait videos that enable the transfer of head motion, face expression and eye gaze. Given a short video of the target actor, we will impose a real-time reenactment algorithm.

Reenactment aims to transfer the motion of a source actor to an image or video of a target actor. Realistic facial expression creation and transformation has been a long- standing problem in computer graphics and computer vision. Thus far popular approaches usually require a driving source or the combination of multiple ones such as capturing a subject's performance and then transferring it to virtual faces. The novelty of the approach lies in the transfer and photo-realistic re-rendering of facial deformations and detail into the target video in a way that the newly-synthesized expressions are made as indistinguishable from a real video as possible.

To achieve this, a dataset will be created of expressions from the target video, then using a webcam or any camera as input feed sources facial traits will be captured and the targets face will be manipulated according to that. For each frame, the environment is considered and ensures that only facial expressions and motions change and everything else remains intact

otherwise the viewer might feel something is out of place. A major challenge is convincing the re- rendering of the synthesized target face into the corresponding video stream. This requires careful consideration of the lighting and shading design which both must correspond to the real- world environment.

III. RELATED WORK

Many types of research have been done on different methods for face recognition, detection and re- enactment here some of the prominent works done in this field are discussed.

A. Offline Re-enactment

Vlasic et al perform facial re-authorization by following a face layout, which is re-rendered under various demeanor parameters over the objective; the mouth inside is straightforwardly replicated from the source video [1]. Picture based offline mouth re-activity was appeared in. Garrido et al. propose a programmed simply picture based way to deal with supplant the whole face. These methodologies only empower self-re-enactment; i.e., when source and target are a similar individual; interestingly, we perform a re-enactment of an alternate objective entertainer.[1] Late work presents virtual naming, an issue like our own; in any case, the technique runs at moderate disconnected rates and depends on conventional teeth intermediary for the mouth inside. Li et al. recover outlines from a database depending on a comparability metric. They utilize optical ow was appearance and speed measure and quest for the k-closest neighbors dependent on timestamps and ow separation. Saragih et al present an ongoing symbol movement framework from a solitary picture.[1] Their methodology depends on inadequate milestone tracking, and the mouth of the source is replicated to the objective utilizing surface twisting.

B. Online Re-enactment

As of late, first online facial re-enactment approaches dependent on RGB-(D) information have been proposed. Kemelmacher-Shlizerman et al empower picture-based puppetry by questioning comparative pictures from a database. They utilize an appearance cost metric and consider pivot rakish distance [2]. While they accomplish noteworthy results, this recovered stream of appearances isn't transiently intelligible. Thies et al show the first online re- enactment framework; be that as it may, they depend on profundity information and utilize nonexclusive teeth intermediary for the mouth district.

C. Viola-Jones Algorithm

The Viola-Jones calculation is a broadly utilized instrument for object discovery [3]. The fundamental property of this calculation is that preparation is moderate, yet identification is quick. This calculation utilizes the Haar premise include channels and Ada support classifier as a modifier, so it doesn't utilize duplications.

The efficiency of the Viola-Jones algorithm can be significantly increased by first generating an integral image [3].

II(y,x) = $\sum_{P=0}^{y} \sum_{q=0}^{x} Y(p, q)$

Detection happens inside a detection window. A minimum and maximum window size are chosen and for each size, sliding step size is chosen.

Feature Extraction Using PCA:

- PCA is utilized to separate highlights from an image of the human face.
- It is utilized as a tool in predictive analysis and explanatory data analysis and is used to transform higher dimension data to lower dimension data.
- PCA gives 72% accuracy.
-

ANN for face Recognition:

In this phase, the data is taken from the images that are simulated from the previously trained ANN. The input will be a vector array from the previous stage. The network is trained with face descriptors as input. The number of Vector will be equal to the number of persons in the database.

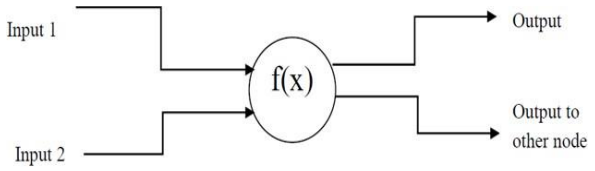

Fig. 3. ANN for Face Recognition

D. Offline RGB Performance Capture

Ongoing disconnected execution catch procedures approach the hard-monocular recreation issue by fitting a blend shape or a multi-straight face model to the info video grouping [4]. Indeed, even geometric _ne-scale surface detail is separated using in-stanza concealing based surface re- enactment. Ichim et al. manufacture a customized face rig from just monocular input. They perform a structure-from- movement recreation of the static head from an explicitly caught video, to which they _t a personality and articulation model. Individual explicit articulations are found out from a preparation arrangement. Suwa-janakorn et al [4]. take in a personality model from an assortment of pictures and track the facial liveliness dependent on a model-to-picture stream field. Shi et al. accomplish amazing outcomes dependent on worldwide vitality improvement of a lot of chose keyframes. Our model-based packaging detailing to recoup on-screen character personalities is like their methodology; however, we use powerful and thick worldwide photometric arrangement, which we implement with client information equal optimization system on the GPU [4]

E. Online RGB-D Performance Capture

Weise et al catch facial exhibitions continuously by fitting a para-metric mix shape model to RGB-D information however they require an expert, custom catch arrangement [4]. The main constant facial execution catch framework based on a ware profundity sensor has been exhibited by Weise et al. Follow up work concentrated on restorative shapes, powerfully adjusting the blend shape premise, non-unbending cross-section distortion, and heartiness against conclusions [4]. These works accomplish great outcomes however depend on profundity information which is regularly inaccessible in most video film.

F. Image-to-Image Translations

A large portion of the ongoing techniques that play out the undertaking of picture to-picture interpretations exploits the intensity of GANs. For the situation that matched information focuses are exhibited, pix2pix [5] performs picture interpretation between two spaces dependent on the misfortune and an ill-disposed misfortune. Notwithstanding when just names between two areas are accessible, Cycle GAN rather misuses the cycle consistency misfortunes between them. All things considered, these techniques are not versatile particularly for picture interpretations on numerous spaces since two sets of generators and discriminator are required for every conceivable area interpretation. As of late, Star GAN proposed to tackle this issue by utilizing just a single generator. Contrasted with earlier techniques, Star GAN empowers various areas interpretation by combining objective space traits with the given picture by linking them channel-wise.

G. Video-to-Video Translations

Several works focus on video-to-video generations. Inspired by Cycle GAN, Recycle GAN [6] translates video contents between two specific domains. In addition to cyclic consistency loss, Recycle GAN imposes
Spatiotemporal constraints between creating realistic results between two seen video domains. Focusing on face-related frameworks, X2Face proposed to synthesize videos based on a learned face representation image extracted from a sequence of source identity videos. Based on driving videos or other conditions such as head poses or audio inputs, the method generates a sequence of driving vectors which in turn move the embedded face image to produce a target video. Another interesting work on video-to-video translations that are based on sparse landmarks is Dyad GAN which generates face expressions in dyadic interactions [6]. The method proposes to produce the video of the interviewer based on the video of the interviewee. The framework consists of two stages, one to generate sketched images of the target domains from the source domain and the other to generate face images based on the sketch.

H. Generative Adversarial Networks (GANs)

GANs are a class of generative models. A GAN as a rule comprises of two contending systems: a generator and a discriminator [6]. The discriminator will likely recognize genuine and created tests while the generator attempts to produce models as practical as conceivable to trick the discriminator. The challenge between the two systems impacts the generator to create increasingly sensible and less foggy outcomes. The first system was produced for picture age from regularly disseminated arbitrary clamors, yet the structure has been received by the network to handle different issues, for example, cGAN and picture to picture interpretations. Because of the prevalence of the system, there are a few GAN-based works, that expand the technique and enhance picture quality just as the security during preparing.

I. Facial manipulation using GANs

Pix2pixHD utilized GANs for high goals picture to-picture interpretation by applying a multi-scale cGAN design and including a perceptual misfortune. GAN proposed a double generator cGAN adapted on feeling activity units, that produces a consideration map. This guide was utilized to interject between the re-enacted and unique pictures, to save the foundation. GAN notation proposed profound facial re- enactment driven by face milestones. It creates pictures logically utilizing a triple consistency misfortune: it initially focuses a picture utilizing tourist spots at that point forms the frontal face. Kim et al. as of late proposed a mixture 3D/profound strategy.[7] They render a recreated 3DMM of a particular subject utilizing a great realistic pipeline. The rendered picture is then handled by a generator arrange, prepared to delineate perspectives regarding each matter to photograph practical pictures. At last, include unraveling was proposed as a method for face control. RSGAN unravels the dormant portrayals of face and hair while FSN proposed an inert space that isolates character and geometric parts, for example, facial posture and demeanor.

J. 3D based methods.

The most punctual swapping strategies required manual contribution. A programmed technique was proposed a couple of years after the fact. All the more as of late, Face2Face moved appearances from source to target face. The move is performed by a 3D morphable face model (3DMM) to the two faces and afterward applying the appearance parts of one face onto the other with care given to inside mouth locales.[13] The re-enactment technique for Suwajanakorn et al. blended the mouth some portion of the face utilizing a reproduced 3D model of (previous president) Obama, guided by face tourist spots, and utilizing a comparative methodology for the face inside as in Face2Face. The appearance of frontal faces was controlled by AverbuchElor et al by moving the mouth inside from source to target picture utilizing 2D wraps and face tourist spots. At long last, Nirkin et al. proposed a face-swapping technique, demonstrating that 3D face shape estimation is pointless for sensible face swaps. Rather, they utilized a 3D face shape as the intermediary. Like us, they proposed a face

division method, thought their work was not starting to finish trainable also, required unique consideration to end.

K. GAN-based methods

GANs [11] were appeared to produce counterfeit pictures with a similar conveyance as an objective space. Albeit effective in creating reasonable appearances, preparing GANs can be shaky and limits their application to low-resolution pictures. Consequent strategies, in any case, improved the dependability of the preparation procedure train GANs utilizing a dynamic multiscale plot, from a low to high picture goals. Cycle GAN proposed a cycle consistency misfortune, permitting preparing of solo nonexclusive changes between different spaces. A CGAN with L1 misfortune was applied by Isola et al. [11] to infer the pix2pix strategy was appeared to create engaging amalgamation results for applications, for example, changing edges to faces.

IV. PROPOSED METHODOLOGY

A. Detect landmarks

Basics of facial landmarks include:

-What facial landmark exactly is and in what way do they work.

-The most effective method to identify and extract facial landmarks from an image with the use of Dlib, OpenCV, and Python. The process of Detecting the facial landmarks is subset of the shape prediction problem. When an input image (and normally an ROI that specifies the object of interest) is given, the shape predictor tries to localize key points of interest along with shape of it.

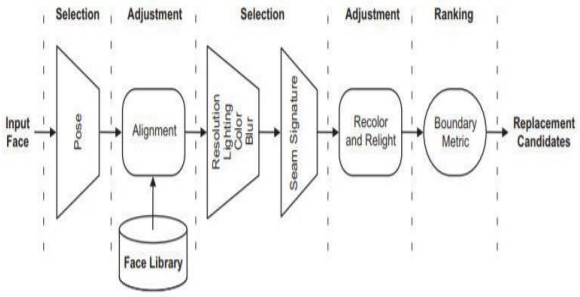

Figure.6. Project Flow

With respect to the facial landmarks, using shape prediction methods our goal will be to the detect important facial structure on the face of the target actor.

Detection of facial landmarks is a two-step process: Step 1: Localizing face in the image.

Step 2: Detection of the key facial structures on the face ROI.

The step one of face detection can be achieved in various different ways. Over the face region, we can then use step 2 for the detection of key facial structure of face region.

There are various types of facial landmark detectors, but these all methods try labelling and localizing some facial region like Mouth, Right eye, Left eye, Nose and Jaw.

B. Resize

The pictures utilized in the investigation are resized in various scales to decide how different sizes influence the acknowledgment procedure. Distinctive picture sizes convey diverse data that is the reason the best picture size should be analyzed in subtleties. The motivation behind picture resizing is to create a lower information size, which rushes the handling time. The resize scale haphazardly shifts from

0.1 to 0.9 worth, which produces diverse picture sizes. shows a case of picture resizing with a size of 0.5. Resizing the picture to a little scale can prompt the loss of numerous significant highlights, particularly if the picture surface is utilized during grouping.

C. GAN (Generative Adversarial Network)

GAN represents Generative Adversarial Network where Generative methods we will produce some likelihood circulation work which turns out to be near the first information which we need to rough Adversarial implies conflict or on the other hand restriction i.e at least two systems will mostly discriminator and generator which will fight among themselves to gain proficiency with the likelihood work. GAN is a profound neural system model contained two neural networks, competing for one against the other. They are neural systems that are prepared in an ill- disposed way to produce information copying some dispersion. The Discriminator is a classifier that decides if the given picture is “genuine" and “fake". The Generator takes an arbitrarily produced commotion vector as info information and input from the Discriminator and creates new pictures that are as near genuine pictures as possible. The Discriminator utilizes the yield of the Generator as preparing data. The Generator gets criticism from the Discriminator. These two models fight" to one another. Every model gets more grounded in the process. The Generator continues making new pictures and remaining its procedure until the Discriminator can never again differentiate between the created pictures and the genuine preparing pictures.

D. CGAN vs DCGAN

Deep convolutional GAN (DCGAN):

The ideal model can learn a multi-modal mapping from inputs to outputs by feeding it with different contextual information.

The architecture of DCGAN (Deep convolutional generative adversarial networks)

The following steps are repeated in training of DCGANs:

- First, the **Generator** creates some new examples.
- The **Discriminator** is trained using real data and generated data.
- After the **Discriminator** has been trained, both models are trained together.
- The **Discriminator**’s weights are frozen, but its gradients are used in the **Generator** model so that the **Generator** can update its weights.

Preparing of DCGANs:

The accompanying advances are rehashed in preparing:

1. The Discriminator is prepared to utilize genuine and counterfeit information and produced information.

2. After the Discriminator has been prepared, the two models are prepared together.

3. First, the Generator makes some new models.

4. The Discriminator's loads are solidified, yet its angles are utilized in the Generator model so the Generator can refresh its loads.

In any case, one of the hindrances of gan is that we don't have control over the yield that Gan is creating that is the reason we are utilizing cgan to deliver the yield we need.
Conditional GANs (CGANs):
Conditional GANs (CGANs) is an extension of the GANs model. CGANs are allowed to generate images that have certain conditions or attributes. Like DCGANs, Conditional GANs also has two components.
1. A Generator (An artist) neural network.
2. A Discriminator (An art critic) neural network.

In conditional GANs (CGANs) including a vector of highlights controls the yield and guides Generator to make sense of what to do. Such a vector of highlights ought to get from a picture which encode the class(like a picture of a lady or a man on the off chance that we are attempting to make countenances of fanciful entertainers) or a lot of explicit qualities we anticipate from the picture (in the event of non-existent on-screen characters, it could be the sort of hair, eyes or appearance). We can consolidate the data into the pictures that will be found out and furthermore into the Z input, which isn't arbitrary any longer. Discriminator's assessment is done not just on the likeness between counterfeit information and unique information yet besides on the correspondence of the phony information picture to its information name (or highlights). We can utilize the equivalent DCGANs and forced a condition on both Generator's and Discriminator's sources of info. The condition ought to be as a one-hot vector adaptation of the digit. This is related to the picture to Generator or Discriminator as genuine or phony.

E. Overall Dataflow Diagram

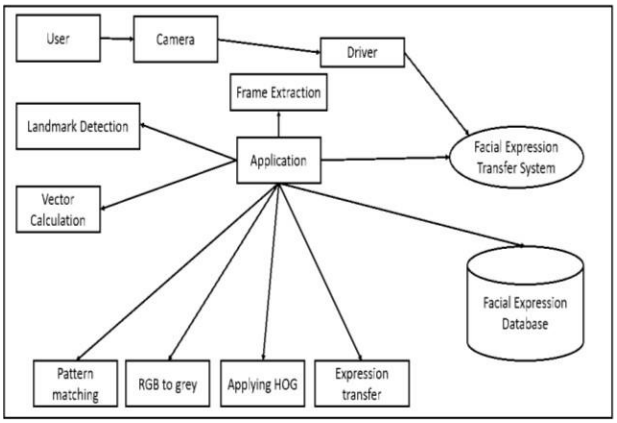

Fig. 9 Dataflow of application

In this dataflow diagram as it can be seen that first image frame is captured and after that, the eyes position is detected after getting the proper positions the position and distance are validated, then face is cropped and resized and then it is rotated and scaled accordingly and then the eigenfaces are computed after all these the distance in the projection space is calculated.

F. Generate Training data(frames)

The preparation and test information is produced by a likelihood dispersion over datasets called the information creating process. We ordinarily make a lot of suppositions referred to altogether as the presumptions. These suppositions are that the models in each dataset are free from one another and that the preparation set and test set are indistinguishably appropriated, drawn from the same likelihood dispersion as one another. This presumption empowers us to portray the information producing process with a likelihood conveyance over a solitary example. The same appropriation is then used to create each train model and each test model. We consider that common fundamental conveyance the information creating circulation, signified information. This probabilistic system and the (i.i.d). presumptions empower us to scientifically consider the connection between preparing mistakes and test blunder. There is a typical presumption that information that is being demonstrated is autonomous and indistinguishably conveyed (i.e.) tests from likelihood dissemination. There is the equivalent fundamental likelihood dissemination for both the preparation and test datasets. Furthermore, each example is autonomous of different examples.

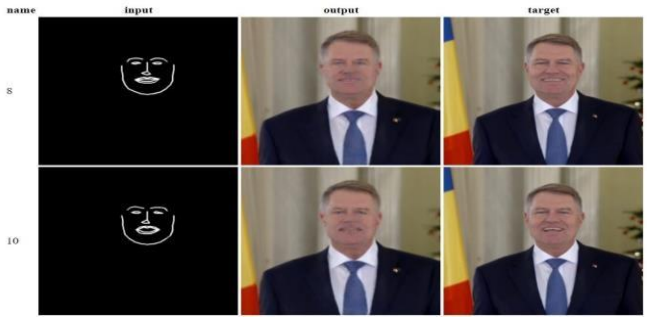

Fig. 11. Frames

V. RESULTS AND DISCUSSION

The below image shows our raw face re-enactment results without background removal. We chose examples of varying ethnicity, pose, and expression. A specifically interesting example can be seen in the rightmost column showing our method's ability to cope with extreme expressions. To show the importance of iterative re-enactment, Fig 4.1 provides re-enactments of the same subject for both small and large angle differences. As evident from the last column for large angle differences, the identity and texture are better preserved using multiple iterations. We report quantitative results, to how we defined the face-swapping problem: we validate how well methods preserve the source subject identity while retaining the same pose and expression of the target subject.

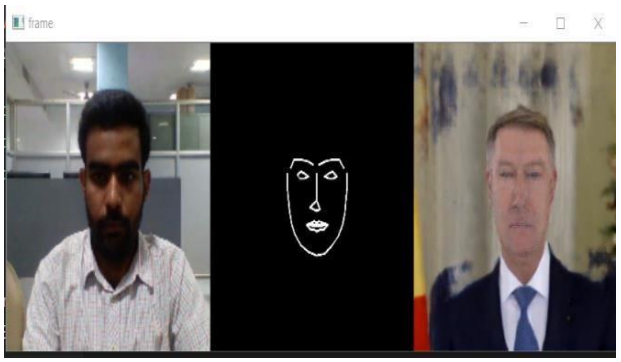

Fig. 12. Result

To this end, we first compare the face-swapping result, Fb, of each frame to its nearest neighbour in a pose from the subject face views. We use the dlib face verification method to compare identities and the structural similarity index method (SSIM) to compare their quality. To measure pose accuracy, we calculate the Euclidean distancebetween thethat our method retainspose and expression much better than its baselines. Note that the human eye is very sensitive to artefacts on faces. This should be reflected in the quality score but those artefacts usually capture only a small part of the image and so the SSIM score does not reflect them well.

VI. CONCLUSION

After using the technologies mentioned in the paper above, we have got our first result that is a video-based re-enactment. In which we record the source video and store it and the target video expression will be changed according to the source video character. There might come changes in the algorithms as the project progresses and also the flow might change as we go for more research on the flow and algorithms for time being we will be using CGAN which currently provides the highest accuracy according to the author of one of the paper still it is to be tested on our level but the results might not differ much in case of accuracy.

We proposed a very flexible methodology for editing facial images according to a target motion defined by a set of facial landmarks. Our methodology can be used for both facial expression/motion transfer, as well as the generation of an image sequence given a single facial image and the sequence of landmarks. We propose a novel way of training such a model to be robust to error accumulation. We demonstrate highly realistic video sequence creation driven by various poses and expressions.

REFERENCES

[1] G. G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape. Ofine deformable face tracking in arbitrary videos. The IEEE International Conference on Computer Vision (ICCV) Workshops, December2015.

[2] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and
C. Theobalt. Real-time expression transfer for facial reenactment. ACM Transactions on Graphics (TOG), 34(6), 2015.

[3] Narayan T. Deshpande and Dr. S. Ravishankar, "Face Detection and Recognition using Viola-Jones algorithm and Fusion of PCA and ANN", in Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 5 (2017) pp. 1173-1189.

[4] Ritu Tiwari, Chandra Prakash Meena, Dhirendra Sharma, and A.Shukla, "Face Recognition using the morphological method" in Indian Institute of Information Technology and Management Gwalior, India April 2009 DOI: 10.1109/IADCC.2009.4809067.

[5] P.Isola, J.-Y.Zhu, T.Zhou,andA.A.Efros., "Image-to-image translation with conditional adversarial networks. Arxiv", 2016

[6] Justus Thies, Michael Zollhöfer, Christian Theobalt, Matthias Nießner "Face2Face: Real-time Face Capture and Reenactment of RGB Videos" in University of Erlangen- Nuremberg and Max-Planck-Onstitute of Infomatics 2015.

[7] Yuval Nirkin, Yosi Keller, Tal Hassener, "FSGAN: Subject Agnostic Face Swapping and Reenactment," arXiv:1908.05932v1, 16 Aug 2019.

[8] Kritaphat Songsri-in, Stefanos Zafeiriou, Imperial College London, University of Oulu," Face Video Generation from a Single Image and Landmarks",arXiv:1904.11521v1, 25 Apr 2019

[9] THIES J., ZOLLHOFER M., STAMMINGER M., THEOBALT C., NIESSNER M.," Face2face: Real-time face capture and reenactment of RGB videos," In Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016), IEEE.

[10] Enrique Sanchez and Michel Valstar. Triple consistency loss for pairing distributions in gan-based face synthesis. arXiv preprint arXiv:1811.03492, 2018.

[11] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Rsgan: face swapping and editing using face and hair representation in latent spaces. arXiv preprint arXiv:1804.03447, 2018.

[12] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, "On face segmentation, face swapping, and face perception. In Automatic Face & Gesture Recognition (FG 2018)," 2018 13th IEEE International Conference on, pages 98–105. IEEE, 2018.