Half-Yearly Progress Report for July-Dec 2021

Data Sheet for M.S Scholars

Name:	**Arti Keshari**
Registration No:	**CS19S008**
Department:	**Computer Science and Engineering**
Date of Joining:	**July 2019**
Specialization / Stream:	**Computer Vision**
Area of Research work:	**Video Generation**
Category of Admission:	HTRA
Guide:	**Prof. Sukhendu Das**
Co-Guide(s):	

Date of GTC meetings:

Description	Event	Date
1st GTC meeting	Research Scholar will have a Mid-Term Review meeting. GTC may recommend on continuation of HTRA.	1.5 years
2nd GTC meeting	Seminar	04/02/2022
3rd GTC meeting	Submission of Synopsis	1 month before thesis submission

Details of Course work

S.No	Course No.	Course Title	Sem/Year	Credits	Grade
Core Courses					
1	CS5691	Pattern Recognition and Machine Learning	01	15	B
2	CS6015	Linear Algebra and Random Processes	01	12	B
3	CS6350	Computer Vision	01	12	S
Elective Courses					
4	CS6730	Probabilistic Graphical Model	02	12	C
5	EE5175	Image Signal Processing	02	12	C
Compulsory Courses / Optional Courses					
6	ID6021	Introduction to Research	01	0	P
7	ID6020	Introduction to Research(Institute Module)	01	0	P

Signature of Scholar

Arti Keshari

Signature of Guide

Signature of Co-Guide 1

Signature of Co-Guide 2

Contents

i) Title of Research Work:

- **Video Generation**

ii) Problem Definition / Research Objectives:

Generative Models (VAE & GAN) are getting very popular nowadays because they can generate unseen data by learning the underlying training data distribution. Image generation models have achieved enormous improvement in generating fine quality fake images. The video generation task has not achieved this level of success because of the high complexity of the video. Due to the existence of temporal dimension, videos have variations in color, speed, initial frame content, intensity, change in content over the frames, etc. Unconditional Video generation is a way to map latent space noise vectors to video samples. The unconditional video generation task takes a randomly sampled Gaussian noise vector as input and returns a video from the distribution learned using training data.

iii) Summary of Work Done before Review (From the date of admission till now)

- First, two-semester was packed with seven courses (including Introduction to Research) and figuring out my research interest. In the third semester, I started reading papers and corresponding hands-on codes. I read about 25+ papers related to Text to Image Generation, Text to Video Generation, Image-to-Image generation, Style transfer GAN, Image to Video Generation, and a few basic concepts.

- Image captioning is a widely studied area in the computer vision field. In which the machine learns how to summarize an image in a single sentence. Text to image generation is the reverse process, in which the machine learns to generate an image using a given caption. This is a less explored research area, so the results are not promising till date. My research objective was to improve existing work by modifying the learning techniques and construct a better model architecture.

- January 2021: I worked to improve performance in the Text to Image generation model and used DFGAN as base paper. Tried to change loss functions first, and experimented with different types of loss function like hinge loss, magp loss, image mismatching adversarial loss, Wasserstein loss with/without singular value clipping, etc. Then tried to incorporate edge maps along with the RGB image to preserve the structure information. However, it could not improve the result.

- February 2021: Since any change in DFGAN was not improving the result, later I made AttnGAN as my base paper to work on. Tried different deep learning techniques to improve the matrix, for example, self-attention, spectral clustering, different image up-sampling technique, incorporated enhancing image resolution method, etc., to get higher resolution image by preserving image quality. Changing the up-sampling method helps to achieve better results than base paper. However, it could not outperform state-of-the-art.

- March/April 2021: Start collaboration with one of my lab-mate and start working on the Text to Video Generation problem. It is a very less explored research area; limited papers have been published to date. We chose MoCoGAN as our base paper and did extensive experiments to map Text and video on common latent space. We experimented with the dimension of the input noise vector, and tried several architectural changes. In addition, to increase the video guidance, we tried to generate optical flow to generate the next video frame by warping technique. In addition, we also tried a frame difference discriminator. Nevertheless, no techniques worked.

- May/June 2021: To solve the Text to video problem, we had read several papers in unconditional video generation, future frame prediction, single image to video generation, optical flow synthesis etc. Slowly we realize first we should work to improve the results of Unconditional Video generation, because multi modality data needs ample data to learn mapping between text and video. After two months of extensive experiments, we submitted a paper in the 32nd British Machine Vision Conference (BMVC) on 25 June 2021. This paper got accepted for publication. I have included the objective of our paper in the Research objective section, and following is the brief of the paper:

- **V3GAN: Decomposing Background, Foreground and Motion for Video Generation**

  - We proposed to decompose the problem of video generation into three subtasks: background, foreground, and motion. In figure1 novelties are: (i) Three-branch generator architecture to decompose foreground, background, and motion without any supervision. (ii) Feature-level masking technique and (iii) Shuffling loss for the discriminator.

  - Figure 1 (left) contains the block diagram of proposed V3GAN architecture. Symbols F, B, and M represent the foreground, background and mask of the generated video. Figure 1 (right) shows the generated background, foreground, mask and final video with no supervision.

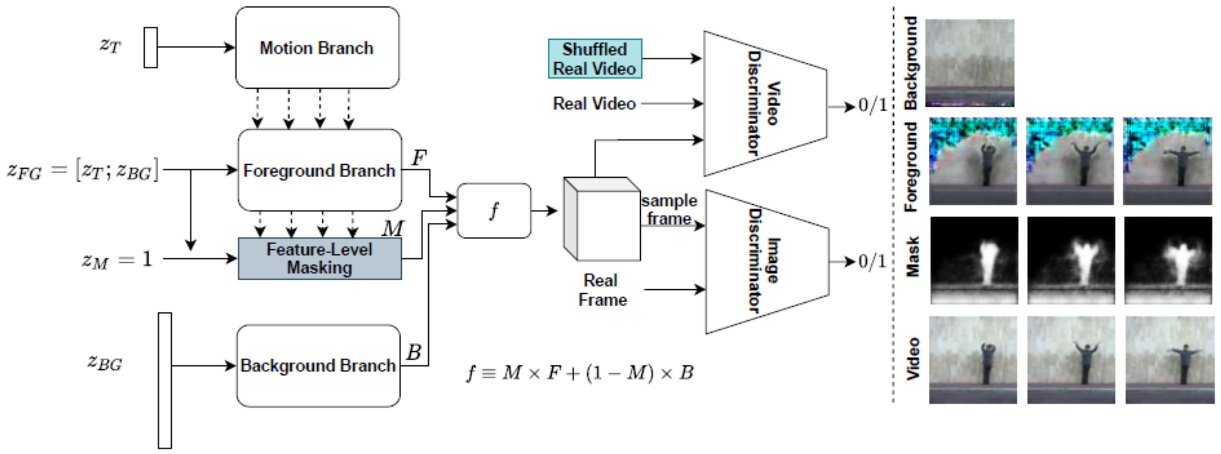


Figure 1: Left: Overview of proposed V3GAN architecture. Right: Illustration of background and foreground, the mask estimated, and the final generated video obtained by combining the three former components.

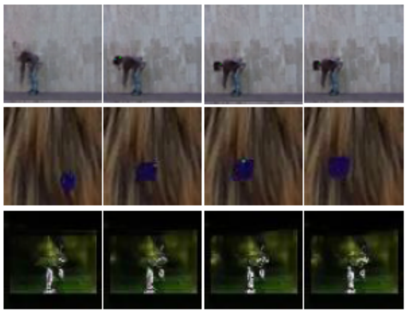
  - **Feature-Level Masking:** This method allows the generator network to learn masking at the feature level as well. It helps to generate sharper masks for separating foreground and background.
  - **Shuffling Loss:** We proposed a novel shuffling loss ($L_{shuffle}$) to achieve better temporal consistency among video frames.

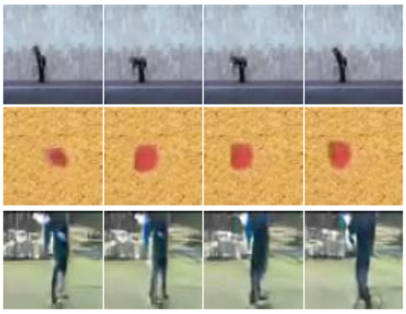
$$L_{shuffle}(X) = E[\log(1 - D_V(sh(X,\alpha)))]$$

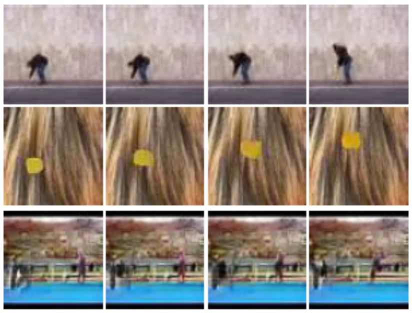
where X is the input real video, D_V is the video discriminator, and α is defined as the fraction of frames that has been shuffled. sh(.) is the shuffling function which takes X and α as input and gives shuffled video.

- **Dataset:**
    - **Shapes Dataset** contains 4000 videos of two shapes, namely circle and square of different colors and sizes moving on 7 different backgrounds. Shapes are moving from left to right, or top to bottom.
    - **Weizmann Action** Dataset contains 93 videos of 9 people performing 10 actions, including jumping-jack and waving-hands. For data augmentation we did horizontal flip.
    - The **UCF101** dataset is commonly used for video action recognition. It includes 13220 videos of 101 different action categories.

- Following are the qualitative results, compared with previous state of the art methods MoCoGAN[2], G3AN[1]:


MoCoGAN


G3AN


V3GAN (Ours)

- Following are the quantitative results, compared with the previous state of the art methods:

Method	**Shapes**	**Weizmann**	**UCF101**		**UCF101 (Stable)**
	(FID↓)	**(FID↓)**	**(FID↓)**	**(IS↑)**	**(FID↓)**
VGAN [4]	-	158.04	115.06	2.94	-
TGAN [3]	-	99.85	110.58	2.74	-
MoCoGAN [2]	144.87†	92.18	104.14	3.06	218.59†
G3AN [1]	-	86.01	91.21	3.62	-
G3AN*	168†	68.19†	86.73†	3.44	102.13†
V3GAN (w/o shuffle)	62.59	64.33	**78.71**	3.84	75.64
V3GAN + shuffle (Ours)	**28.07**	**62.65**	80.18	**3.88**	**74.36**

Table 1: Quantitative comparison with SOTA methods using FID metric for Shapes, Action, UCF101 including stabilized UCF101 datasets. † indicates that the values are obtained by training the official codes provided by the authors.

Failure Case

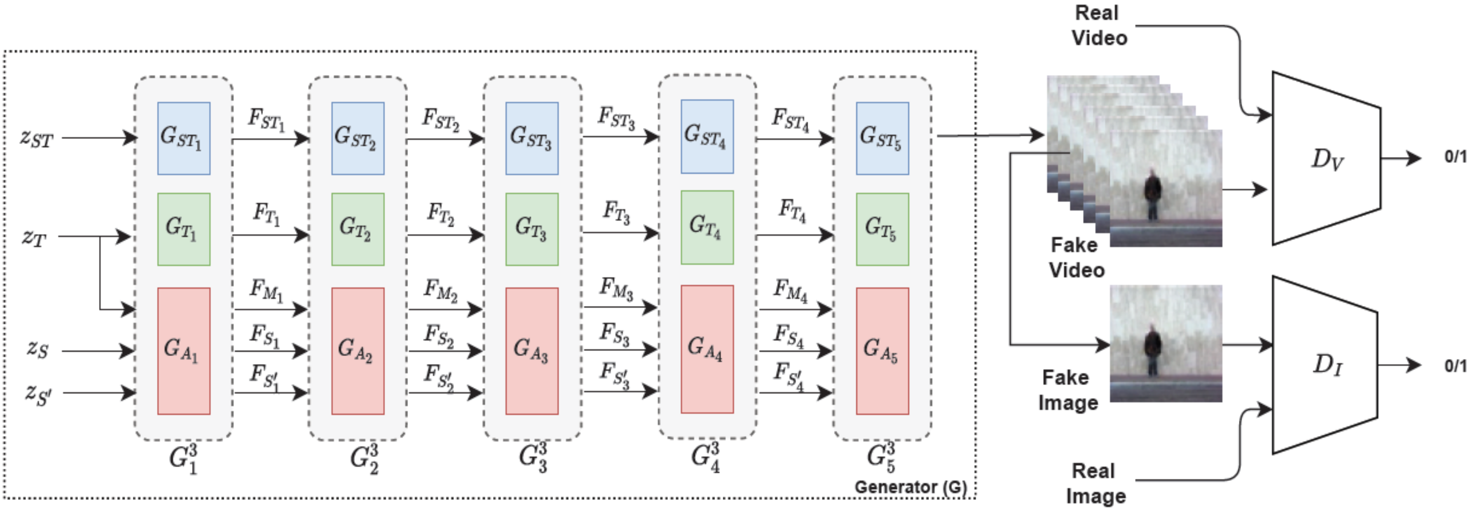
iv) Work Done During Review (Even Semester 2021), in chronological order:

- July/August 2021: To solve the task of video generation, we make the existing video generation network wider by splitting the spatial stream into two parallel identical branches learning complementary feature representations. We also propose a novel adaptive masking layer to facilitate the learning of complementary features. Our design choice helps us to achieve an improvement of 11.35 on FID metric compared to state-of-the-art method for UvA-NEMO dataset. Wrote a paper titled "G3AN++: Exploring Wide GANs with complementary Feature Learning for Video Generation" in collaboration with Ph.D. scholar, Sonam Gupta and submitted in the 12th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2021). It got accepted for publication.

- **G3AN++: Exploring Wide GANs with complementary Feature Learning for Video Generation**

    - Summary of the main contributions of the above work:
      • We propose a novel wider Generative Adversarial network, G3AN++, which learns a richer feature representation while maintaining the ability to control motion and appearance of the generated videos.
      • We propose a novel adaptive masking layer to facilitate the learning of complementary features by the identical branches used for modeling the appearance stream.
      • Extensive quantitative and qualitative experimentation on the benchmark datasets shows that the proposed method outperforms the state-of-the-art (SOTA) methods by a significant margin.

    - Following are the overall architecture of our proposed model G3AN++ (Figure 2) and detailed architecture of G_A module (Figure 4).



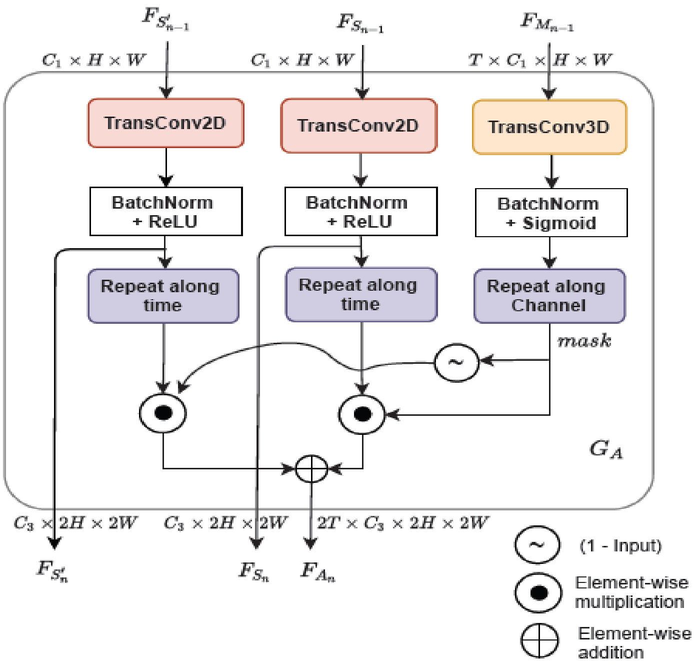
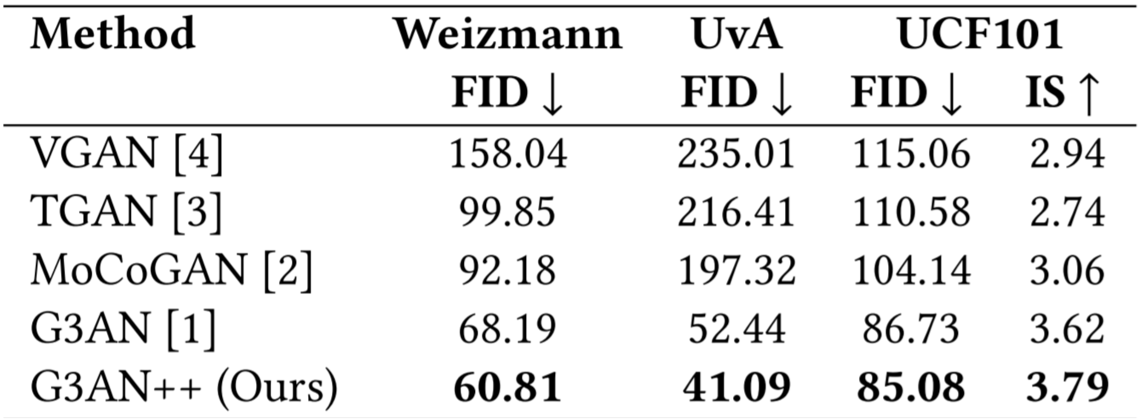


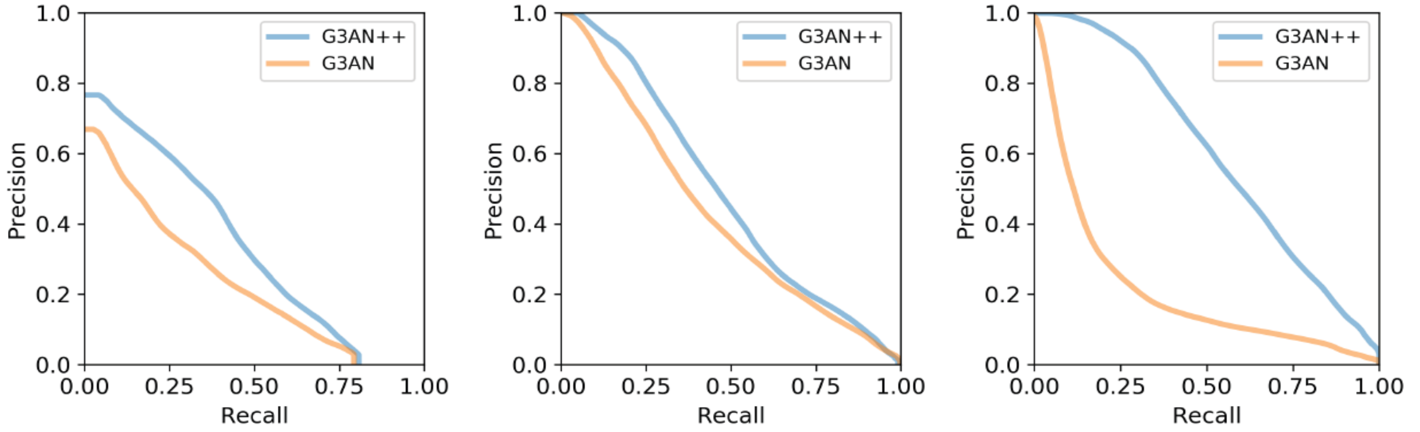
Figure 4: Proposed Dual branch appearance module with adaptive masking layer, within Appearance stream (G_A) of G3AN++ (see Fig. 3). Subscripts S_n and S'_n correspond to the two parallel branches which learn complementary visual features with the help of the proposed adaptive masking layer marked by subscript M.

- Quantitative Results can be seen in Table 1, Figure 5 and Table 2. Table 3 contains quantitative results for ablation study to prove the importance of each component of our network.



Method	Weizmann	UvA	UCF101	
	FID ↓	FID ↓	FID ↓	IS ↑
VGAN [4]	158.04	235.01	115.06	2.94
TGAN [3]	99.85	216.41	110.58	2.74
MoCoGAN [2]	92.18	197.32	104.14	3.06
G3AN [1]	68.19	52.44	86.73	3.62
G3AN++ (Ours)	**60.81**	**41.09**	**85.08**	**3.79**

Table 1: Quantitative results of our proposed G3AN++ method compared with the SOTA methods, on Weizmann, UCF101 and UvA-Nemo datasets. The best results are shown in bold.

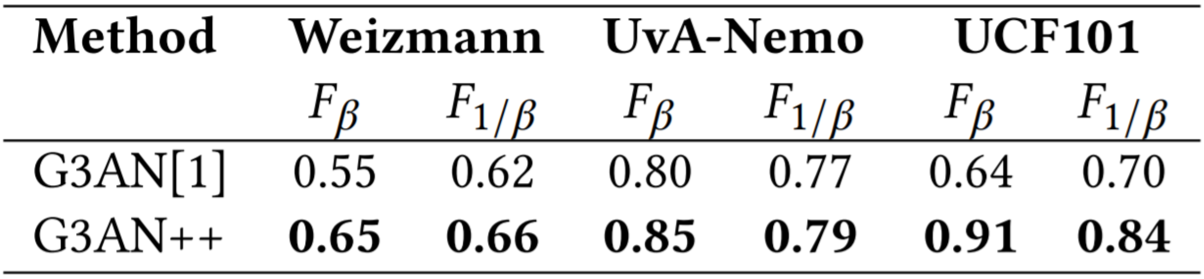


(a)

(b)

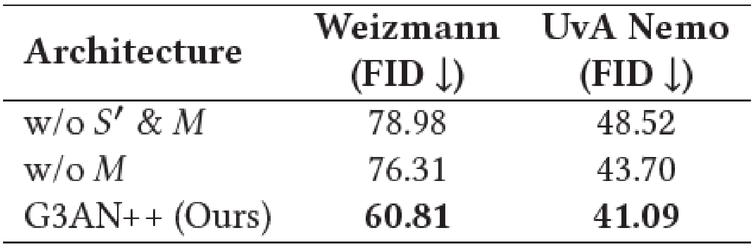
(c)

Figure 5: Precision-Recall curve on (a) Weizmann, (b) UVA-Nemo, and (c) UCF101 datasets for G3AN and G3AN++ (Our proposed) methods.



Method	Weizmann		UvA-Nemo		UCF101	
	F_β	$F_{1/\beta}$	F_β	$F_{1/\beta}$	F_β	$F_{1/\beta}$
G3AN[1]	0.55	0.62	0.80	0.77	0.64	0.70
G3AN++	**0.65**	**0.66**	**0.85**	**0.79**	**0.91**	**0.84**

Table 2: F-scores for Weizmann, UVA-Nemo and UCF101 datasets. The values are reported for $\beta = 8$.



Architecture	Weizmann (FID ↓)	UvA Nemo (FID ↓)
w/o S' & M	78.98	48.52
w/o M	76.31	43.70
G3AN++ (Ours)	**60.81**	**41.09**

Table 3: Ablation study: Importance of individual components of G3AN++ model.

- Qualitative Results can be seen in the following figures. Figure 6, 7, 8 contains results for the Weizmann action dataset, UvA NEMO dataset and UCF 101 dataset.

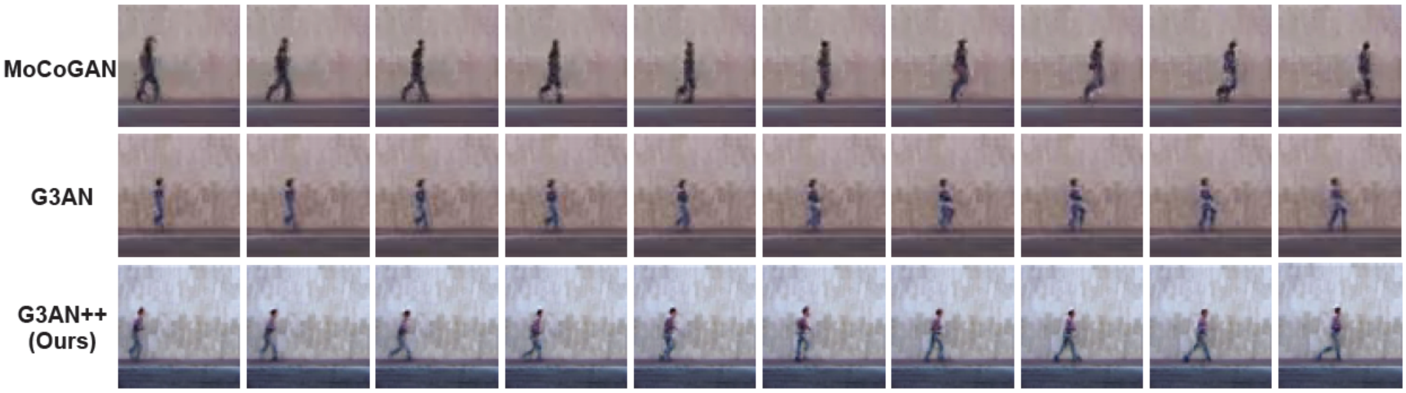


Figure 6: Comparison of the performance of G3AN++ on Weizmann dataset. The first row demonstrates a sample generated by MoCoGAN [27], second row demonstrates the samples generated by G3AN [29] and the last row demonstrates the result generated by our proposed method G3AN++ (figure best viewed in color).

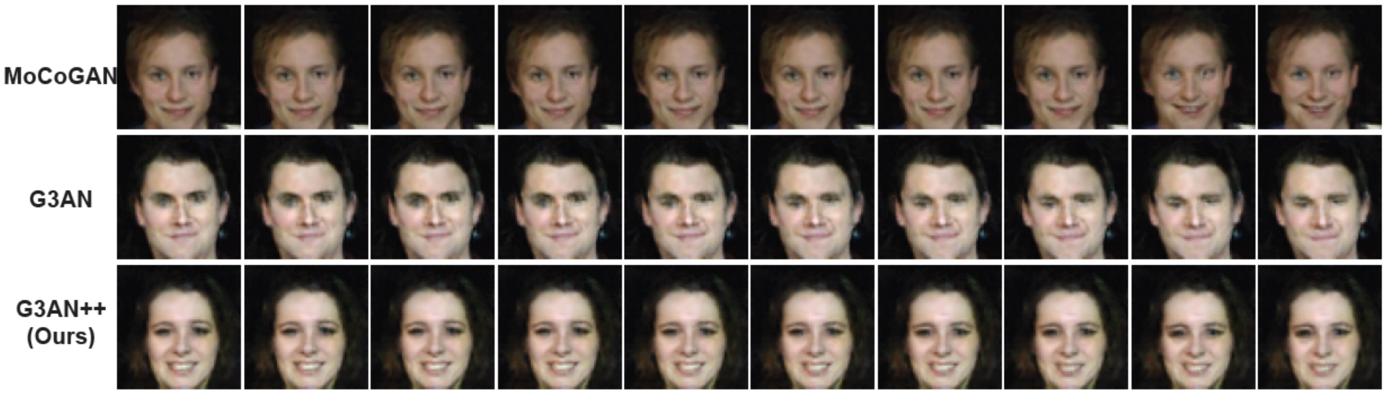


Figure 7: Comparison of the performance of G3AN++ on UVA-Nemo dataset. The first row showcases a sample generated by MoCoGAN [27], second row showcases the samples generated by G3AN [29] and the last row showcases the result generated by out proposed G3AN++ method (figure best viewed in color).

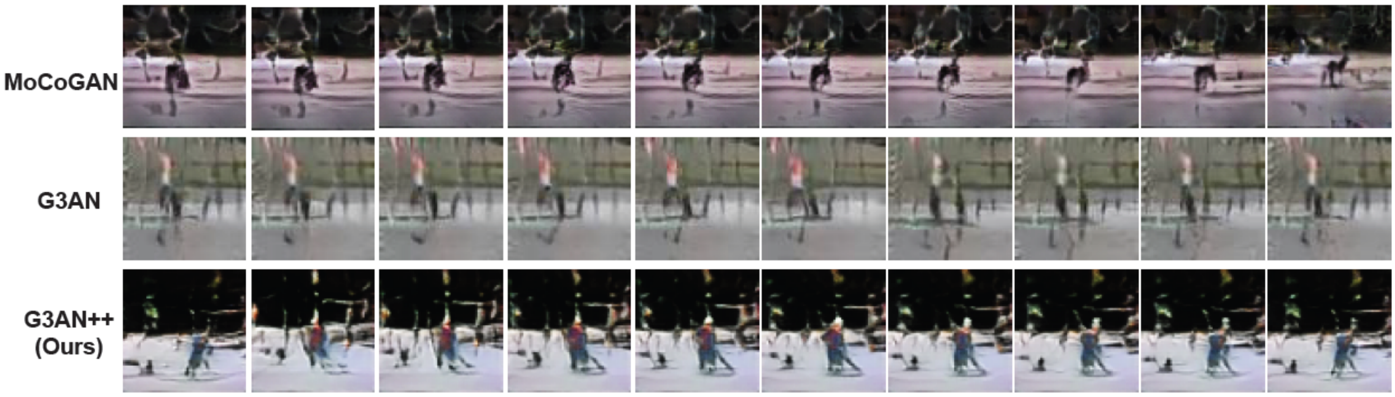


Figure 8: Comparison of the performance of G3AN++ on UCF101 dataset. The first row illustrates a sample generated by MoCoGAN [27], second row illustrates the samples generated by G3AN [29] and the last row illustrates the result generated by our proposed G3AN++ method (figure best viewed in color).

- August-September 2021: Solved the video generation task by a novel hybrid generative framework, TransGAN, which uses a combination of a recurrent generator and 3D convolution based discriminator. We further introduce a novel Transpose Convolutional LSTM (TransConvLSTM) unit that functions as the building block for the recurrent generator. We achieved an improvement of 18.42 on the FID metric compared to the state-of-the-art method for MUG facial expression dataset. Wrote a paper titled "TransGAN: Recurrent GAN using Transpose Convolutional LSTM for Video Generation" and submitted it to the Winter Conference on Applications of Computer Vision (WACV 2022). But it got rejected.

- Prepared rebuttal response for the submitted paper in BMVC also does extra experiments to prove our points.

- October 2021: Read research papers related to recurrent GANs in order to improve the flaws of WACV rejected papers. Also did some experiments to see if the model can give comparable results even with a lower number of deep layers.

- BMVC paper got accepted, submitted a camera ready version, and prepared a presentation slide with the co-author.

- November-December 2021: Right now, I am reading recent papers published in video/image generation problems. Few of the recent works are able to do pose control, translation control and camera angle control for image generative models. I wanted to extend these ideas to controlled video generation.

- Also recently prepared a powerpoint presentation for ICVGIP with co-author and gave the presentation.

v) Issues affecting Research Progress, if any

- Since this area is less explored and needs a high computational machine, it is tedious to find a correct path.
- Working with high resolution video and images takes days to train the model, which makes it hard to switch to high resolution data.
- Due to the pandemic I stayed at home for about 1.5 year, which highly affected my research progress.

vi) Future Plans, with proposed timeline

- Planning to revise the rejected paper and submit it in another upcoming conference.
- I am looking forward to making a controllable GAN for video generation tasks, using the latest techniques of deep learning.
- Use local video data (captured in house), as input for video generation/prediction.

vii) Visible Research Output:

(a) Full Paper(s) Published in Conference Proceedings

[1] "V3GAN: Decomposing Background, Foreground and Motion for Video Generation", Arti Keshari, Sonam Gupta, Sukhendu Das. In the 32nd British Machine Vision Conference (BMVC 2021) Online 22nd - 25th November 2021 [Poster]

[2] "G3AN++: Exploring WideGANs with Complementary Feature Learning for Video Generation", Sonam Gupta, Arti Keshari, Sukhendu Das. In the 12th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2021), December 2021, IIT Jodhpur, India [Oral] DOI: 10.1145/3490035.3490282

(b) Seminars/Workshops/Conferences/Exchange Programmes attended and Papers Presented

[1] Presented the paper titled "G3AN++: Exploring WideGANs with Complementary Feature Learning for Video Generation" at ICVGIP 2021 Conference, on 22nd December.
[2] Attended BMVC 2021 as an author and participated in a poster presentation doubt clearing session.
[3] Attended International Conference on Image Processing (ICIP) 2020 in virtual mode.
[4] Attended Center for Computational Brain Research (CCBR) 2019.

References:

[1]Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3AN: Disentangling appearance and motion for video generation. In Proceedings of the IEEE Conference on CVPR, pages 5264–5273, 2020.
[2] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In Proceedings of the IEEE Conference on CVPR, pages 1526–1535, 2018.
[3] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal Generative Adversarial Nets with Singular Value Clipping. In Proceedings of the IEEE ICCV, pages 2830–2839, 2017.
[4] Vondrick, Carl, Hamed Pirsiavash, and Antonio Torralba. "Generating videos with scene dynamics." NIPS(2016): 613-621.