Half-Yearly Progress Report for Jan-May 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:

Co-Guide(s): **Prof. Sukhendu Das**

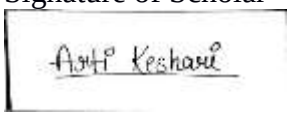
Date of GTC meetings:

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	
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
6	CS6910	Fundamental of Deep Learning	02	12	D
Compulsory Courses / Optional Courses					
7	ID6021	Introduction to Research	01	0	P
8	ID6020	Introduction to Research(Institute Module)	01	0	P

Signature of Scholar



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 underlining 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 temporal dimension, videos show variations like color, speed, initial frame content, intensity, change in content over the frames, etc. Unconditional Video generation is a way to map latent space noise vector to video sample. The unconditional Video Generation task takes Gaussian noise as input and returns a sampled Video from the distribution learned by 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 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 imaginary image using a given caption. This is a less explored research area, so the results are not promising to date. My research objective is to improve existing work by modifying the techniques and adding newer architecture.

- I ran codes to regenerate the current results. Also, learn how to debug codes.

iv) Work Done During Review (Odd Semester 2021)

Following are the work done in chronological order:

- January 2021: I worked to improve performance in the Text to Image generation model using 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 map along with the RGB image to preserve the structure information. However, it could not improve the result.
- February 2021: Again, I choose AttnGAN as my base paper to work on. Tried different levels of changes to improve the evaluation matrix. I used the following methods: self-attention, spectral clustering, changed image up-sampling technique, incorporate 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 experiment with the dimension of the input noise vector, tried several architectural changes. In addition, to increase the video guidance, we tried to generate optical flow to generate next video frame by warping technique. In addition, we also tried frame difference discriminator. Nevertheless, no techniques worked.
- May/June 2021: To solve 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 we should work to improve the results of Unconditional Video generation. After two months of extensive experiments, we submit a paper in the 32nd British Machine Vision Conference (BMVC) on 25 June 2021. I have included the objective of our paper in the Research objective section, and the Following is the brief of the paper:

- 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. Figure1 (right) shows the generated background, foreground, mask and final video with no supervision.

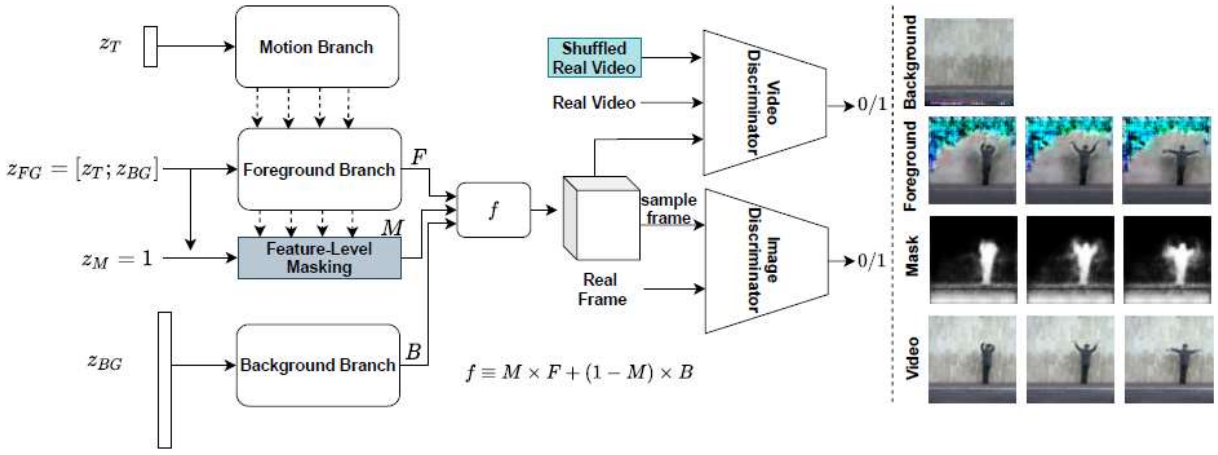
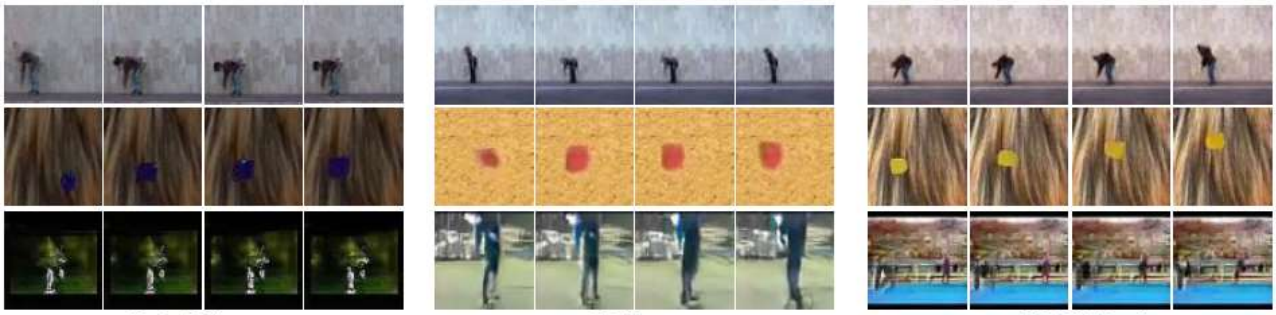


Figure 1: Left: Overview of proposed V3GAN architecture. Right: Illustration of background and foreground, the mask estimated, and the final generated video obtained by combining the three former components.

- **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 to achieve better temporal consistency among video frames.

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

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



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

Method	Shapes (FID↓)	Weizmann (FID↓)	UCF101 (FID↓)	UCF101 (Stable) (FID↓)
VGAN [23]	-	158.04	115.06	-
TGAN [19]	-	99.85	110.58	-
MoCoGAN [22]	144.87†	92.18	104.14	218.59†
G3AN [26]	168†	68.19†	86.73†	102.13†
V3GAN (w/o shuffle)	62.59	64.33	**78.71**	75.64
V3GAN + shuffle (Ours)	**28.07**	**62.65**	80.18	**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.

v) Issues affecting Research Progress, if any

- Since the area is less explored and needs a high computational machine, it is tedious to find a correct path.

vi) Future Plans, with proposed timeline

- We (my collaborator and me) wanted to extend this work and target the next coming conference.
- I wanted to explore possibilities in the de-shuffling image using GAN since no work has been done for such a problem statement.