

Supplementary Material: V3GAN: Decomposing Background, Foreground and Motion for Video Generation

Arti Keshari
cs19s008@cse.iitm.ac.in
Sonam Gupta
cs18d005@cse.iitm.ac.in
Sukhendu Das
sdas@iitm.ac.in

Visualization and Perception Lab
Dept. of CSE
Indian Institute of Technology, Madras
Chennai, India

1 Overview

The supplementary is organized as follows. In section 2, we provide details about the network architecture of V3GAN. In Section 3, we show qualitative examples for Conditional video generation using class labels and Longer sequence (32 frames) generation for Weizmann dataset. In section 4, we present additional ablation studies on the proposed shuffling loss. We also show the architecture diagrams of the ablation studies performed in Section 4.3 of the main paper. Lastly, in section 5, we show more samples of generated videos along with foreground, background and mask on Shapes and UCF101 datasets. We also show some failure cases on Weizmann dataset.

For rest of the supplementary, we follow the notations given below :
 z_{BG} : Background noise vector, z_T : Motion noise vector sampled from Standard Normal distribution, z_{FG} : foreground noise vector constructed by concatenating z_{BG} and z_T . For illustration purpose, only 10 frames of each video are displayed in a row for all figures. Unconditionally generated videos from V3GAN can be found at link: [examples](#).

2 Network Architecture

We discuss the details of the proposed model. V3GAN generator consists of three branches, namely, motion V_T , foreground V_{FG} , and background V_{BG} . The foreground branch is augmented with the proposed feature-level masking layer V_M . V3GAN maps $1 \times 1 \times 1$ noise vectors to a sequence of 16 RGB frames of resolution 64×64 . In Table 2, we present the number of output channels along with the shape of the output features at each stage for every branch in the network. The details of the masking layer has also been shown in a separate column.

	Motion Branch		Foreground Branch		Masking Layer		Background Branch	
Layers	Ch	TxHxW	Ch	TxHxW	Ch	TxHxW	Ch	TxHxW
Input	10	1x1x1	138	1x1x1	1	1x1x1	128	1x1x1
1	512	4x1x1	512	4x4x4	1	4x4x4	512	1x4x4
2	512	8x1x1	512	8x8x8	1	8x8x8	512	1x8x8
3	256	16x1x1	256	16x16x16	1	16x16x16	256	1x16x16
4	128	16x1x1	128	16x32x32	1	16x32x32	128	1x32x32
5	-	-	64	16x64x64	1	16x64x64	64	1x64x64
6	-	-	3	16x64x64	-	-	3	1x64x64

Table 1: Output dimension of each branch/layer in V3GAN. Ch specifies the number of output channels.



Figure 1: Unconditionally generated videos with 32 frames each. Each row represents a video sequence generated by randomly sampling z_M and z_{BG} from standard Gaussian distribution. 10 alternate frames have been picked starting from 3rd frame due to space constraint.

3 Additional Qualitative Study

3.1 Generating Longer Video Sequences

We slightly modify V3GAN to generate 32 frame sequences to investigate the generative capabilities of the network. We observe that even with longer sequences our model shows promising results. However, as expected the spatial as well as temporal quality is degraded. Ghost person (eighth row), person without head (second row), person with no action (seventh row), etc, appears frequently as shown in Figure 1. The FID value 120.35 confirms the observed degradation in the quality of the generated videos.

3.2 Conditional Video Generation

We study the effect of the conditioning the model on class labels for Weizmann dataset. The class information is given to the generator by concatenating one hot vector representation of the category label with temporal noise z_T . For the discriminator, we reshape and repeat the class label representation so that it can be concatenated as a channel with the input video. Similar approach has been followed for Image Discriminator. Figure 2 shows the results obtained by feeding random noise vectors with class information of Weizmann dataset to the generator. It can be seen that V3GAN can successfully generates the samples corresponding to each category of the dataset.

3.3 Latent Space Exploration

To understand the effect of latent vectors z_{BG} and z_T , while generating videos, we use the following set-up. We sample a random noise vector z_{BG} with three different z_T and vice versa. We observe that for constant background noise, the background remain fixed and foreground depends on both z_{BG} and z_T . Figure 3, 4, 5, shows the latent space exploration for Weizmann, Shapes and UCF101 datasets, respectively.

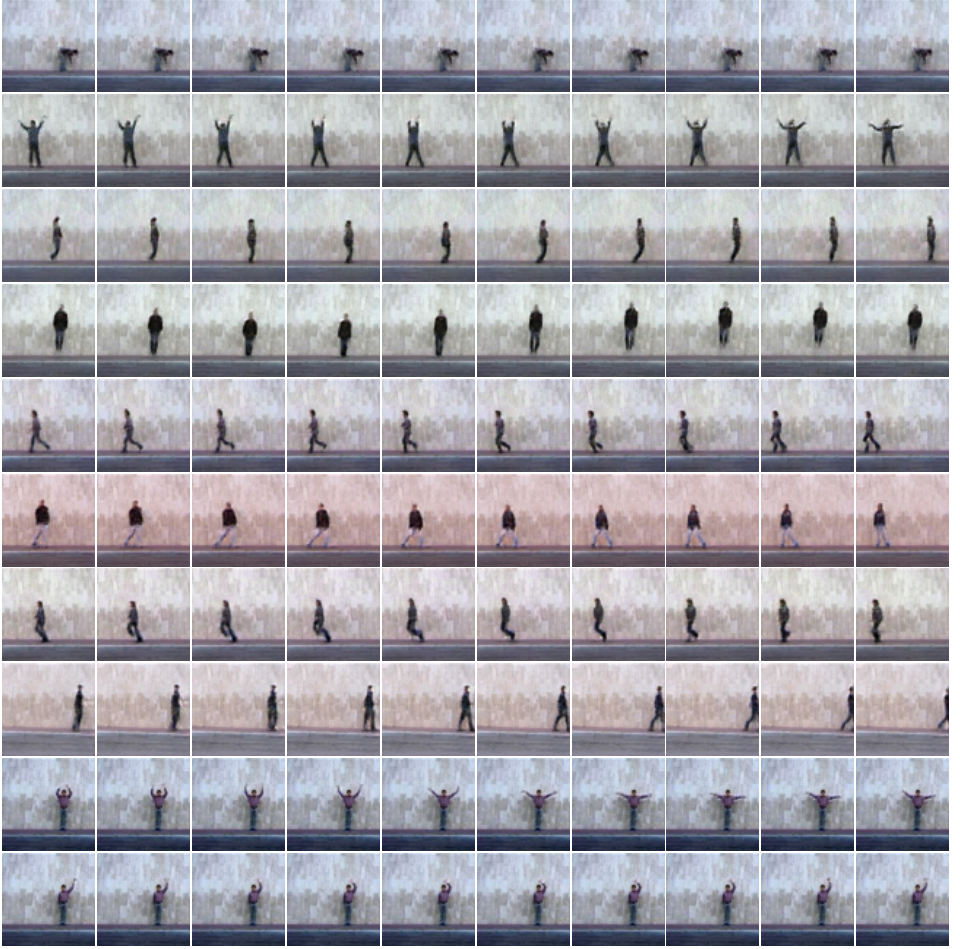


Figure 2: Qualitative results for Conditional video generation from V3GAN. Each row represents a video generated by combining randomly sampled z_M and z_{BG} with one hot vector representation of class label on Weizmann dataset. Class label used, top to bottom: bending, jumping jack, jump, parallel jump, running, side, skip, walking, two hand waving, one hand waving.



Figure 3: Generated samples for Latent space exploration on Weizmann dataset. We sample 3 different z_{BG} and 3 z_T . z_{BG} is kept fixed for three consecutive rows, and z_T is fixed for every $(3n + 1)^{th}$ row, where $n \in 0, 1, 2$. Each row represents a video sequence.

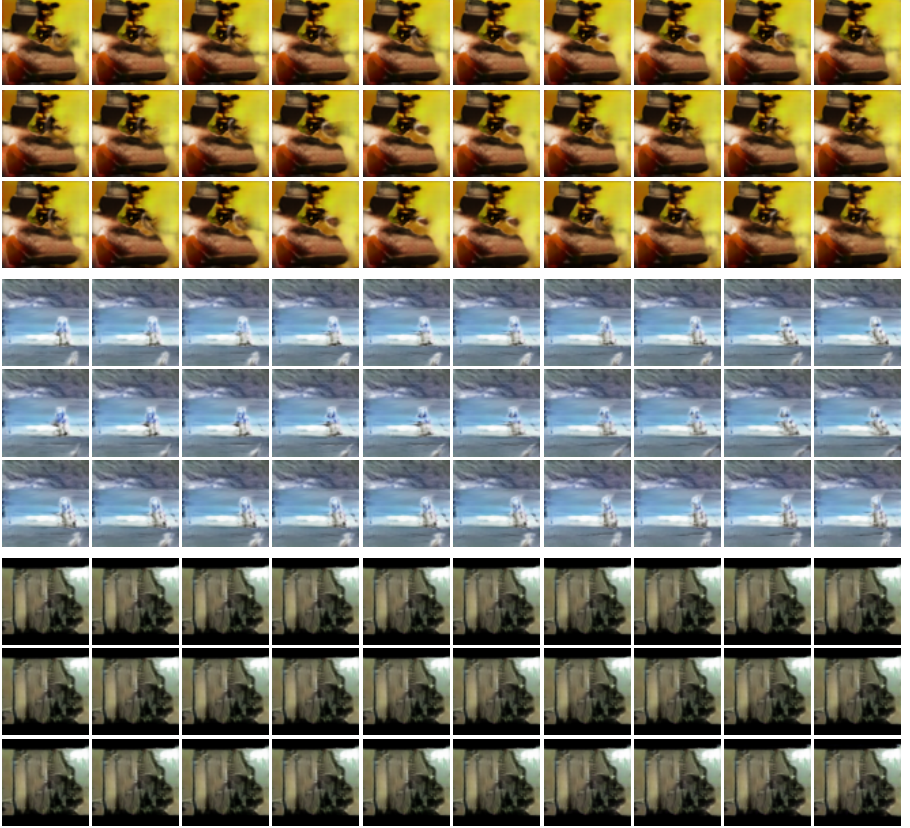


Figure 4: Generated samples for Latent space exploration on UCF101 dataset. We sample 3 different z_{BG} and 3 z_T . z_{BG} is kept fixed for three consecutive rows, and z_T is fixed for every $(3n+1)^{th}$ row, where $n \in 0, 1, 2$. Each row represents a video sequence. First three rows seems to be Playing Daf, Next three rows shows surfing, last three rows corresponds to cliff diving.

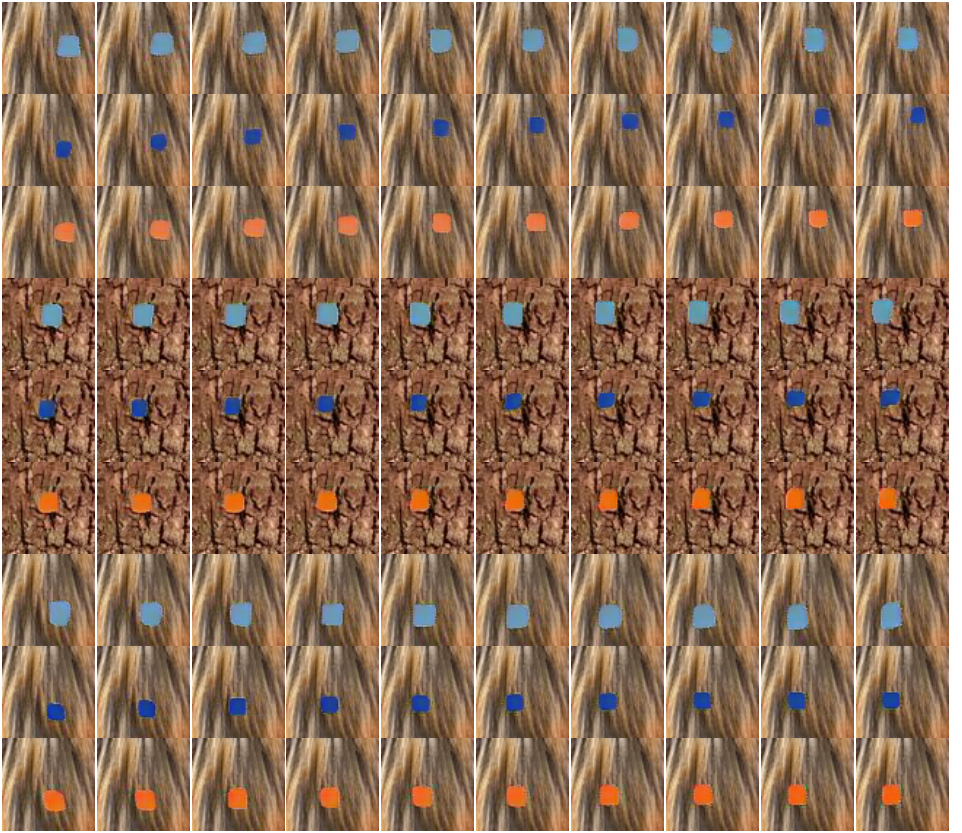


Figure 5: Generated samples for Latent space exploration on Shapes dataset. We sample 3 different z_{BG} and 3 z_T . z_{BG} is kept fixed for three consecutive rows, and z_T is fixed for every $(3n+1)^{th}$ row, where $n \in 0, 1, 2$. Each row represents a video sequence.

4 Additional Ablation Study

4.1 Architecture diagram for ablation studies

In figure 6, we show the individual architectures for various ablations performed in Table 2 of the main paper. Notice how the last masking layer is retained in the configuration (a), (c) and how the entire masking is removed in (d). Such a design is chosen so that we can explicitly examine the effect of the proposed feature level masking.

4.2 Effect of Gamma

To understand the weightage required for shuffling loss, we introduce the scaling parameter γ in video discriminator as shown in equation 1. We train our network with different values of γ like $[10, 1, 0.1, 0.25, 0.5, 0.75, 0.01]$. However, the best results are obtained for $\gamma = 1$.

$$F_V = E[\log D_V(x)] + E[\log(1 - D_V(G(z_{BG}, z_T)))] + \gamma L_{shuffle}(x) \quad (1)$$

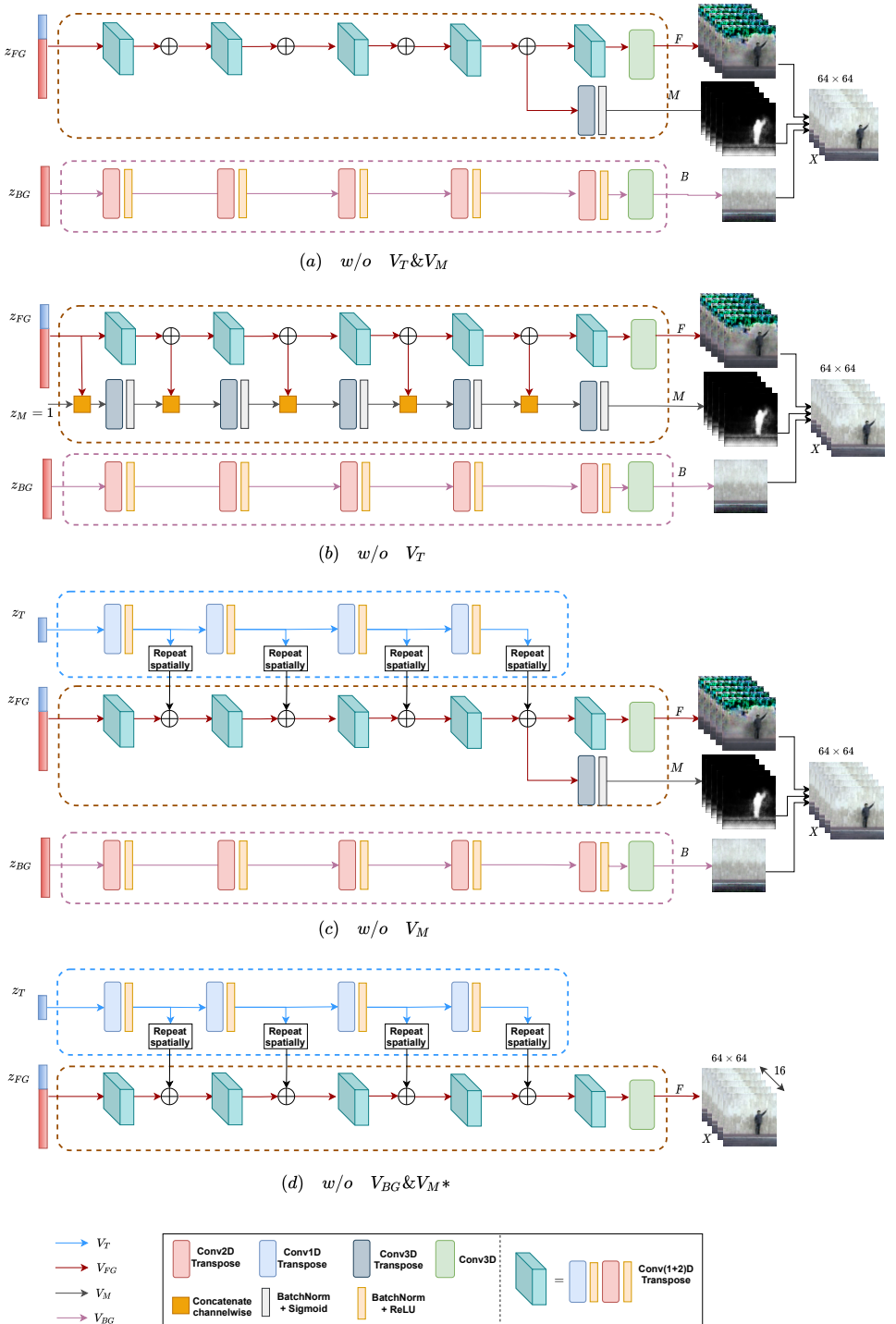


Figure 6: Architecture diagram for different ablation studies performed in the Main paper.

5 More Qualitative Results

5.1 Shapes Dataset and Generated samples on Shapes Dataset

For Shapes dataset, we added a randomly sampled background out of seven textured backgrounds, for each video. Those backgrounds are shown in figure 7. The samples generated from Shapes dataset are shown in figure 8.



Figure 7: Seven background textures chosen for generating Shapes dataset.

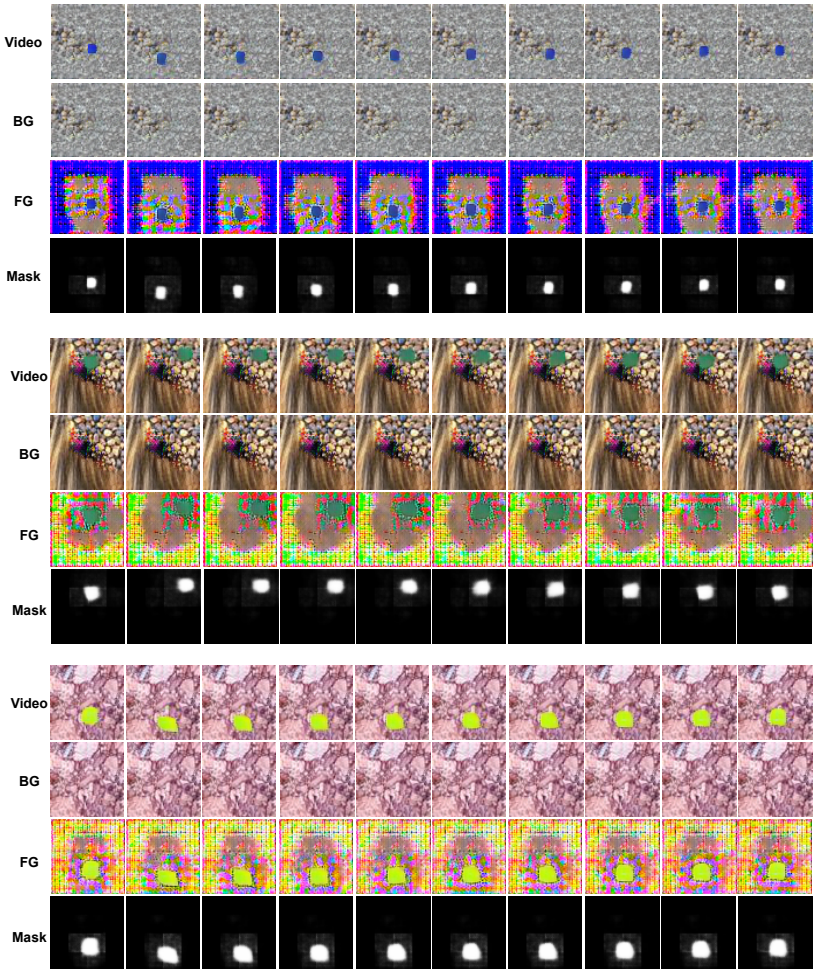


Figure 8: Generated video samples for Shapes dataset, along with corresponding background, foreground and mask generated from V3GAN.

5.2 Generated Samples for UCF101 Dataset

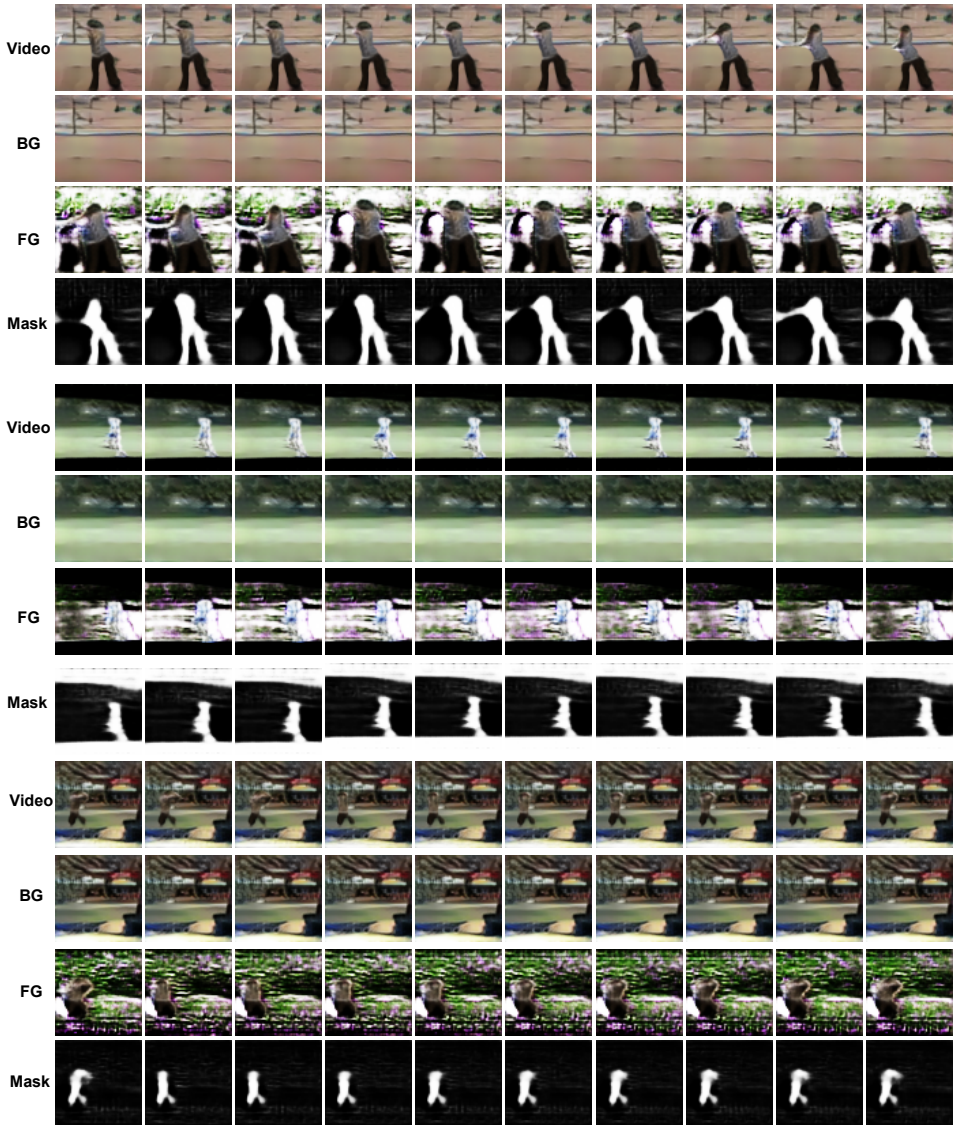


Figure 9: Generated video samples for UCF101 dataset, along with corresponding background, foreground and mask generated from V3GAN.

5.3 Some Failure Cases for Weizmann Dataset



Figure 10: Failure cases in Weizmann Dataset.