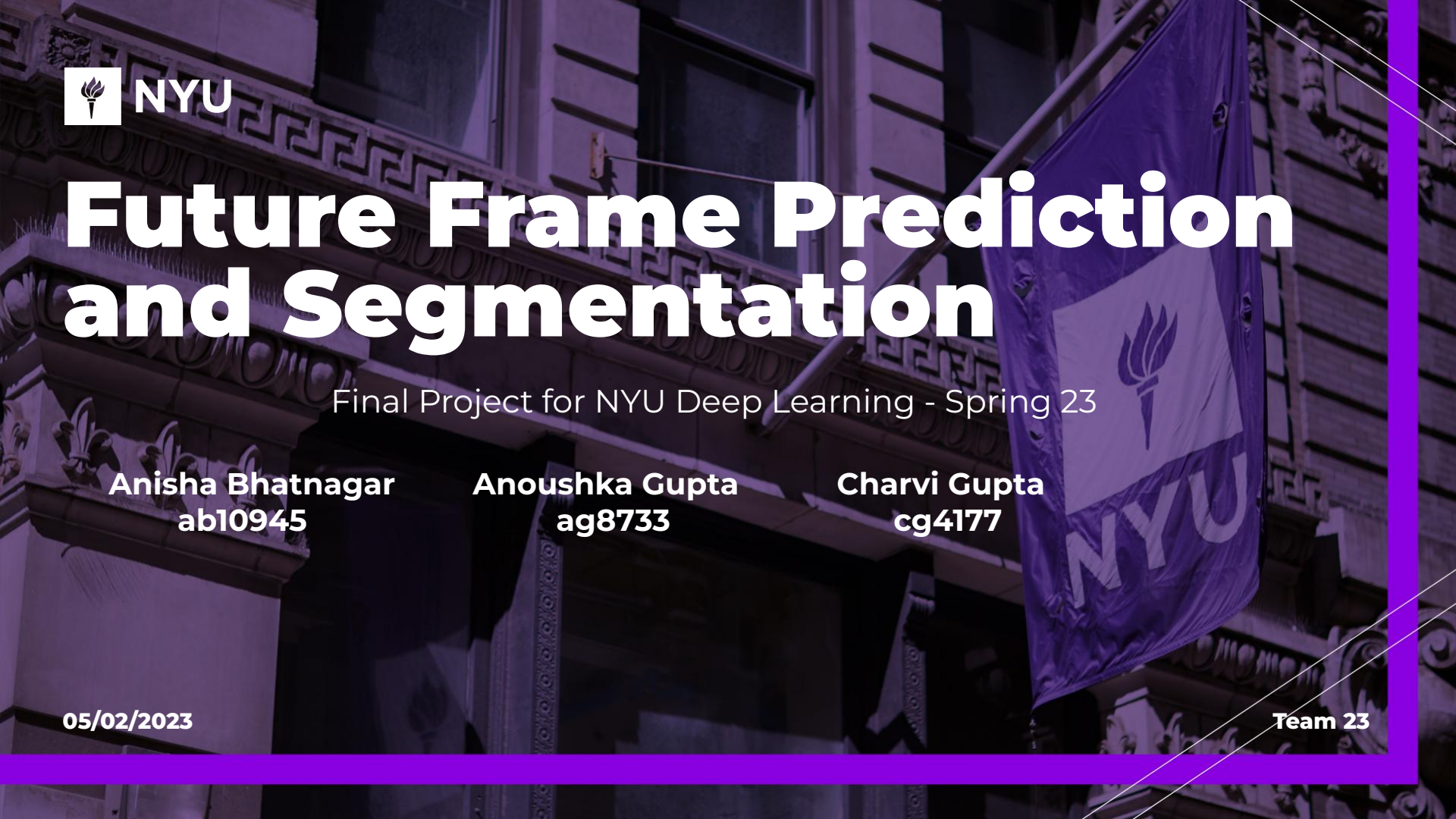


NYU

Future Frame Prediction and Segmentation

Final Project for NYU Deep Learning - Spring 23

Anisha Bhatnagar
ab10945

Anoushka Gupta
ag8733

Charvi Gupta
cg4177

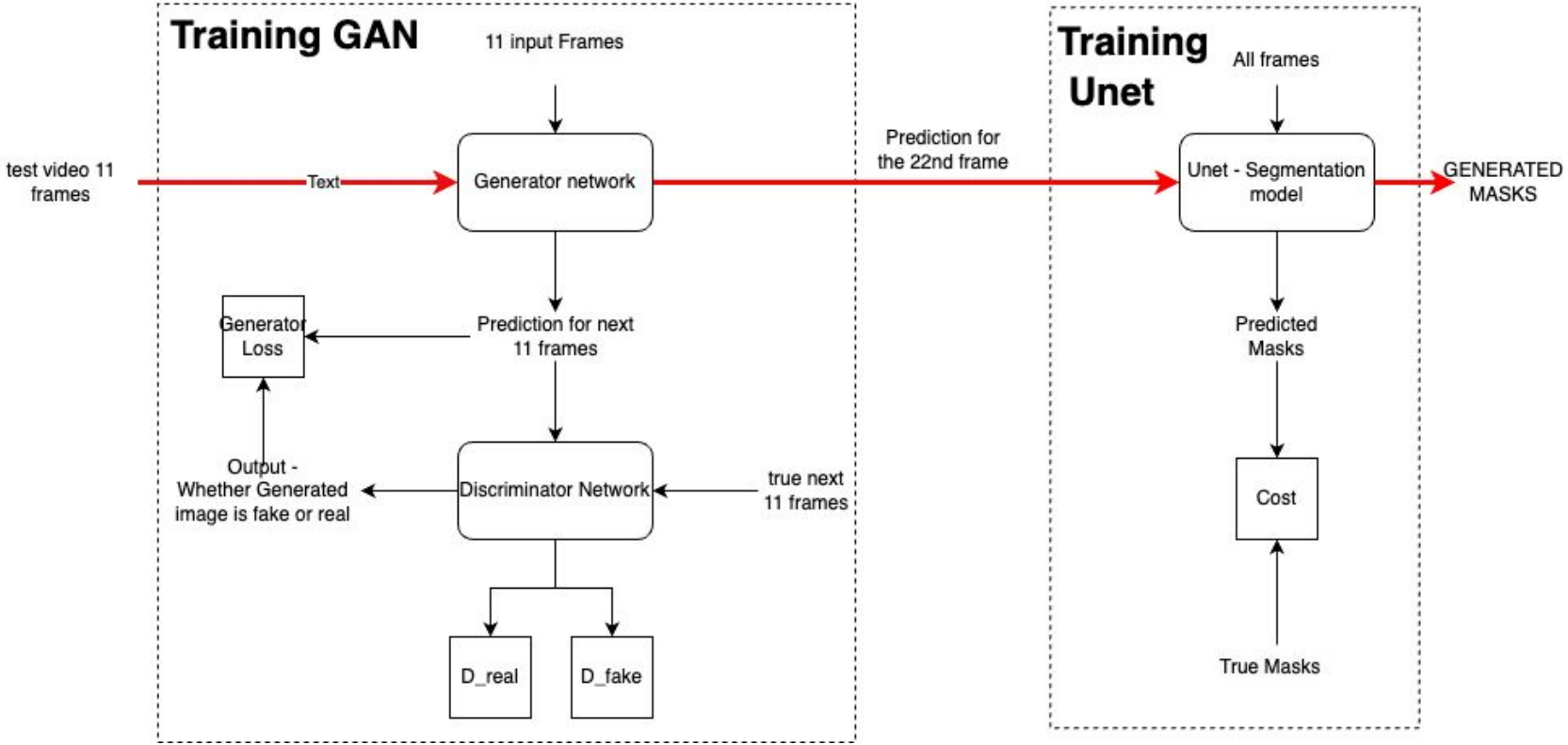
05/02/2023

Team 23

Problem Statement And Literature Review

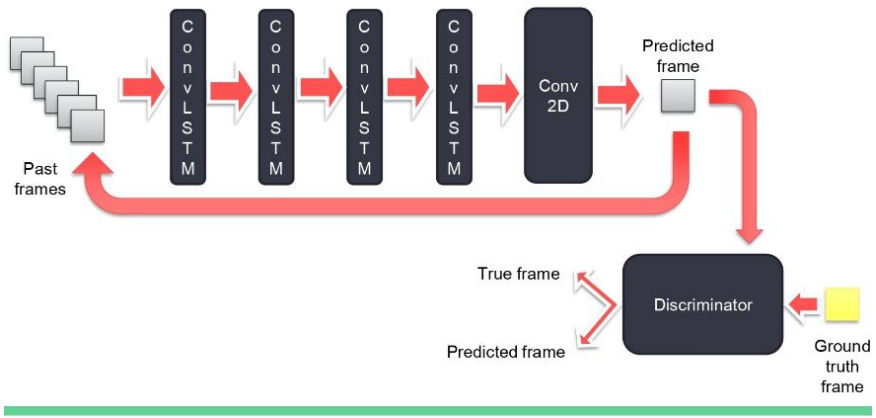
- **Problem Statement** - Using the first 11 frames of a video predict the segmentation mask of the last (22nd) frame.
- **ConvLSTMs** (introduced by Shi et al (2015)) have been used for sequence to sequence predictions as they can effectively capture the spatio temporal features. (other common techniques make use of convolutional AutoEncoders)
- **U-nets** (introduced by Ronneberger et al (2015)) perform well on segmentation related tasks due to the skip connections and the contracting and expansive networks. The contracting (downsampling) network easily learns feature maps and identifies the objects in an image. The expansive(upsampling) network takes the feature map generated by the contracting network and generates a segmentation mask by using the skip connections.

Solution Approach



Future Frame Prediction - Model

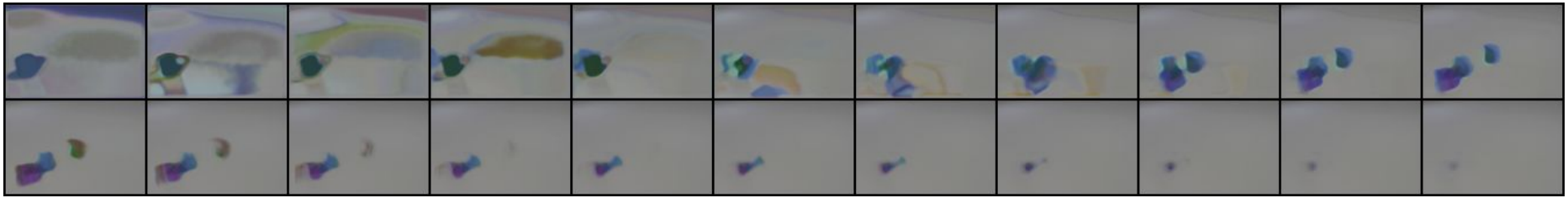
- A GAN was used to perform frame prediction. The Generator model used had 4 convLSTM layers followed by one Conv2D head.
- The 1st 11 frames were used as input into the model. The predicted frames were fed back into the model to continue prediction for the next 10 frames.
- BCE Loss was used to train the discriminator, a combination of l1_l2 loss and BCE loss was used to train the generator.
- Training was performed using 2000 videos out of the unlabeled set for 50 epochs on v100.



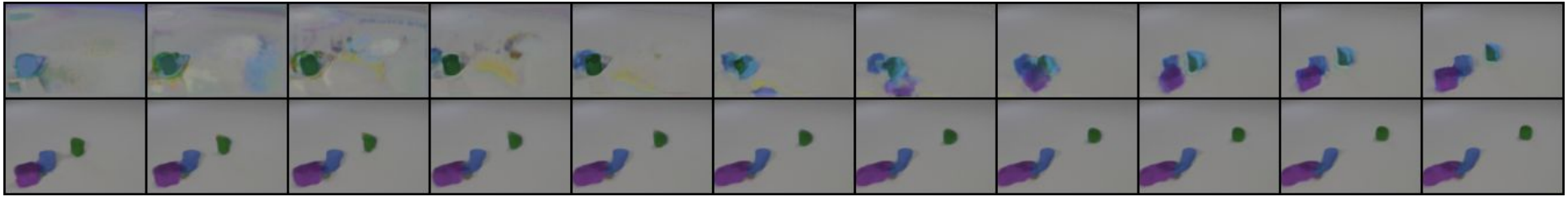
(Image taken from Vineeth S.)

Visualization of intermediate results

After 2000 steps

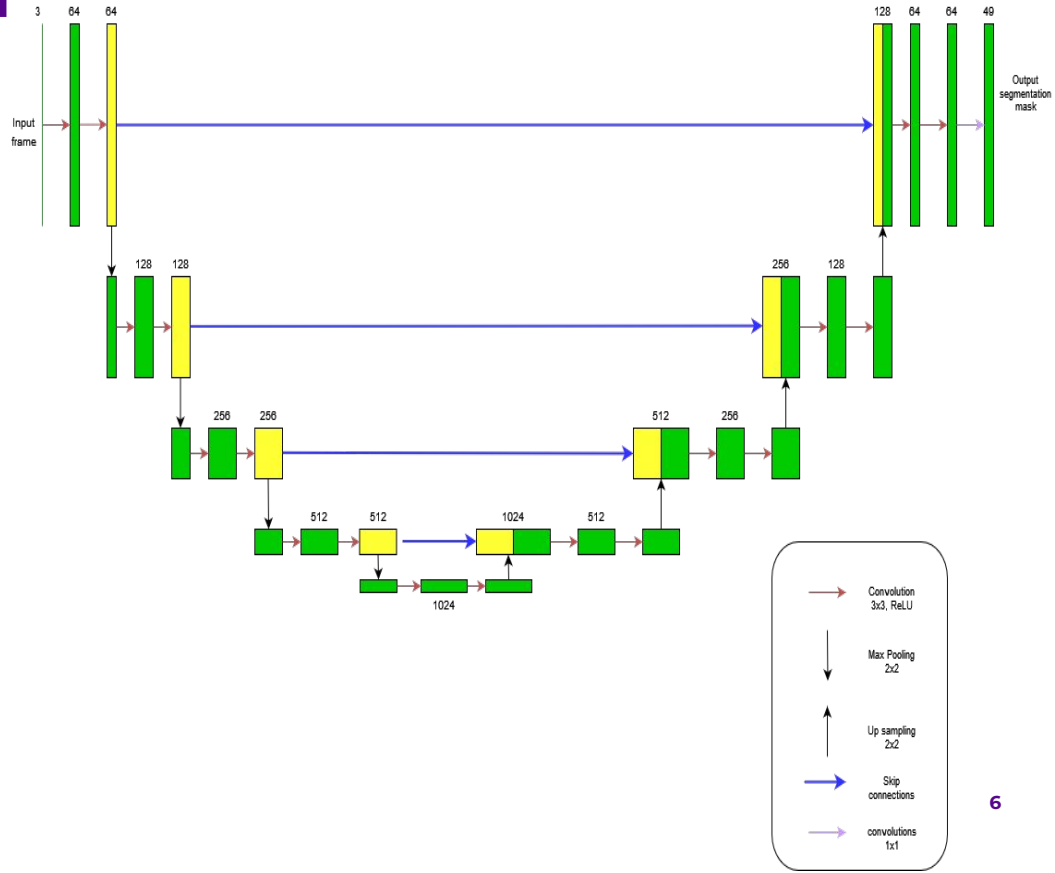


After 25000 steps



Segmentation Model

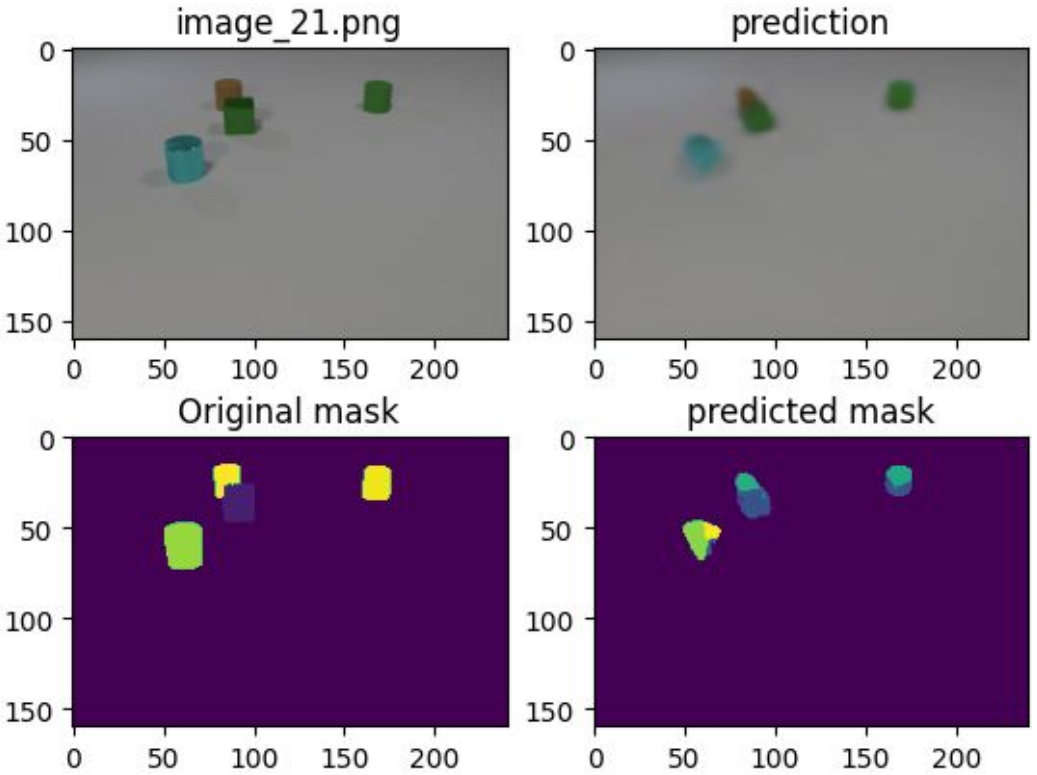
The segmentation model was based on the Unet model which gave an output of 49 x 160 x 240 which is the one hot encoding matrix for each class in the mask.
The model was trained for 10 epochs using CrossEntropyLoss as the loss function and a Adam Optimizer.
The 22nd frame which was predicted by the ConvLSTM model was fed into the Unet model to generate the final masks



NYU

Results

On the validation dataset, which had 1000 videos our model achieved a Jaccard index score of **0.2437**



NYU

Thank you

Charvi Gupta (cg4177@nyu.edu),
Anoushka Gupta (ag8733@nyu.edu),
Anisha Bhatnagar (ab10945@nyu.edu)

NYU