DS-GA 1008 Final Competition Report

Xinran Liao
Center for Data Science
New York University
xl4703@nyu.edu

Hailie Nguyen
Center for Data Science
New York University
hln2020@nyu.edu

Jack Savage
Center for Data Science
New York University
jes9858@nyu.edu

1 Introduction

In the rapidly evolving field of computer vision, the ability to accurately predict and segment future frames in video sequences presents both significant challenges and opportunities for advancements in various applications, from autonomous driving to video surveillance. This report details the efforts of Team 14 in the DS-GA 1008 Final Competition, where our objective was to leverage deep learning models to generate the semantic segmentation mask of the last frame based on the first 11 frames of video sequences.

2 Literature Review

2.1 FutureGAN

FutureGAN, introduced by Aigner and Körner [1], represents a significant advancement in the field of video prediction. The model employs a generative adversarial network (GAN) architecture that leverages spatio-temporal 3D convolutions to predict future frames of video sequences based on past frames. Unlike traditional approaches that may suffer from blurry predictions due to deterministic strategies, FutureGAN effectively captures the inherent uncertainty of video frame prediction through its adversarial training approach. The architecture builds upon the principles of progressively growing GANs (PGGANs), which were originally designed to enhance the quality and stability of generated images. By extending this concept to video, FutureGAN achieves not only high-quality frame predictions but also maintains sharpness and reduces the tendency toward mode collapse, a common issue in GAN training [2].

2.2 U-Net

The U-Net architecture, initially developed for biomedical image segmentation by Ronneberger et al. [3], has been widely adopted for various image-to-image tasks. U-Net features a symmetric encoder-decoder structure that emphasizes precise localization, enabling the model to excel in tasks requiring detailed contextual information from images. The architecture's effectiveness is largely attributed to its use of skip connections, which allow the network to propagate context information directly across the network. In the domain of video frame prediction and segmentation, U-Net can be adapted to handle temporal sequences by incorporating modifications such as recurrent or convolutional LSTM layers, making it suitable for dynamic scene understanding and prediction [4].

3 Methodologies

The dataset provided for this competition includes 1,000 labeled training videos ("train"), 1,000 labeled validation videos ("val"), and 13,000 unlabeled videos ("unlabeled"), each consisting of 22 frames. Our task at hand is to generate semantic segmentation masks for the 22^{nd} frames of the "hidden" dataset consisting of 5,000 videos, each containing the first 11 frames. Our approach

involves two main stages: first, predicting the 22nd frame using the first 11 frames of each video; and second, generating a segmentation mask for the predicted frame.

To address the video prediction challenge, we employed the *FutureGAN* model, which utilizes spatio-temporal 3D convolutions to anticipate future frames. This choice was inspired by the work of Sandra Aigner and Marco Körner[1], who demonstrated the model's effectiveness in generating high-quality predictions by progressively growing GANs.

For the image segmentation task, we implemented the *U-Net* architecture, renowned for its efficiency in pixel-wise classification and its ability to handle data with a high variance in object scales.

3.1 Image Prediction with FutureGAN

We trained *FutureGAN* with a final resolution of 128 x 128px on both the "train" dataset and the "unlabeled" dataset. Unfortunately, due to time and resource constraints, we had to stop training *FutureGAN* on the "unlabeled" dataset at the final checkpoint for the 32 x 32px resolution.

The 22nd frame of each video in the hidden dataset is then resized to 240x160 and given as input for U-Net to generate image segmentation masks.

FutureGAN requires the number of frames for the entire sequence to be available at test time. While this condition is satisfied for the "val" dataset, we only have access to the first 11 frames in the "hidden" dataset. As a result, we duplicated the frames in the "hidden" video folders such that each folder contains 22 frames, the first 11 frames corresponding to the 11 frames on which the prediction is conditioned.

3.2 Image Segmentation with U-Net

We trained the *U-Net* on each frame of the 1,000 labeled training examples and validated the model on the 1,000 labeled validation examples. We built a custom dataset object to load the videos sequentially (to avoid RAM crashes) and to pair them with their masks. We trained for 50 epochs with both validation train loss either decreasing or remaining constant. Even still, when inspecting the results visually we found that earlier epochs' masks looked better on the unlabeled dataset. For our final predictions, we used the checkpoint from the seventh epoch.

4 Results

Our evaluation primarily focused on two metrics for each model used in the video frame prediction and image segmentation tasks.

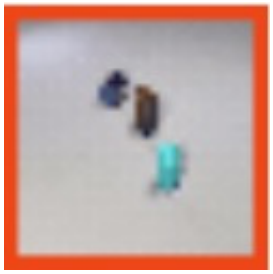
4.1 FutureGAN

For the *FutureGAN* model, Mean Squared Error (MSE) is employed, a metric that measures the average squared difference between the estimated values and what is estimated, providing a clear indicator of prediction precision. Peak Signal-to-Noise Ratio (PSNR) is also used, which assesses the ratio of the maximum possible power of a signal to the power of corrupting noise, indicating image clarity and quality. Lower MSE values and higher PSNR values reflect more accurate and clear predictions.

Table 1 illustrates our validation results for*FutureGAN* trained on the "train" and "unlabeled" dataset, evaluated at their final training resolution (128x128 for the model trained on the "train" dataset, and 32x32 for the model trained on the "unlabeled" dataset).

It is evident from the frame predictions, along with the MSE and PSNR, that the model performs better with a larger training dataset. This improvement can be attributed to the model's enhanced ability to generalize from a more diverse and comprehensive dataset, which increases its robustness when dealing with unseen data.

(a) Trained on 1k “train” videos (128x128)



(b) Trained on 13k “unlabeled” videos (32x32)

Figure 1: 22nd frames predicted by FutureGAN given the first 11 frames of the same validation video

	Train (1k)	Unlabeled (13k)
Average MSE	0.0153	0.0069
Average PSNR	24.6909	28.4933

Table 1: FutureGAN results on "val" dataset

	Train (1k)	Val (1k)
Average BCE	0.015	0.024
Average IOU	0.524	0.467

Table 2: U-Net results

4.2 U-Net

For the *U-Net* model, Binary Cross Entropy (BCE) quantifies the pixel-wise classification error between the predicted masks and the actual masks, directly reflecting the accuracy of segmentation outputs. Intersection over Union (IOU) is another critical metric, measuring the overlap between predicted and ground truth masks, thereby providing a direct assessment of segmentation effectiveness. Unlike BCE loss, IOU is a directly interpretable metric. Qualitatively, when looking at our results 2, we can see that the model segments objects well, but struggles with identifying the correct class. Assuming that our model creates nearly perfect masks for all objects, an IOU of approximately 50% corresponds to a classification accuracy of 50%. Unfortunately, this performance did not generalize well to the predicted frames of the FutureGAN model, as they were often distorted. Interestingly, while the background seems similar between the original frames and the generated frames, *U-Net* predicted one of the object classes for the backdrop.

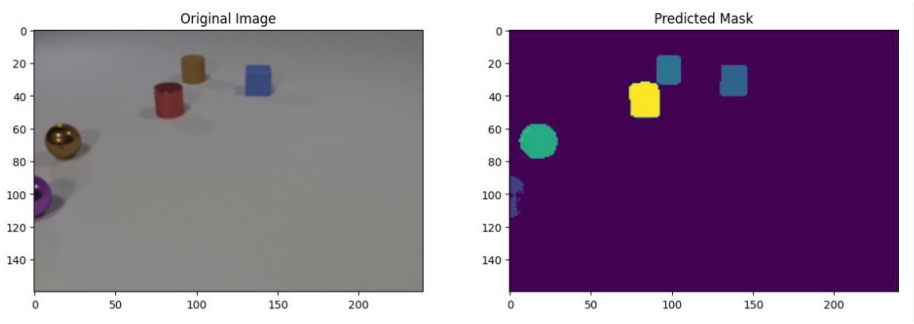


Figure 2: Prediction on “val” video by U-Net trained on 1k “train” videos

5 Future Work

Our experiments with FutureGAN and U-Net have yielded promising results, demonstrating the capability of these models in handling image prediction and segmentation tasks respectively. Based on our findings, we outline several directions for future research:

- **Higher Resolution Training:** Unfortunately, due to time and resource constraints, our current FutureGAN model was trained up to 32x32px on the "unlabeled" dataset. Future experiments could explore the effects of training at higher resolutions, up to 256x256px. This could potentially enhance the detail and accuracy of the generated frames and segmented

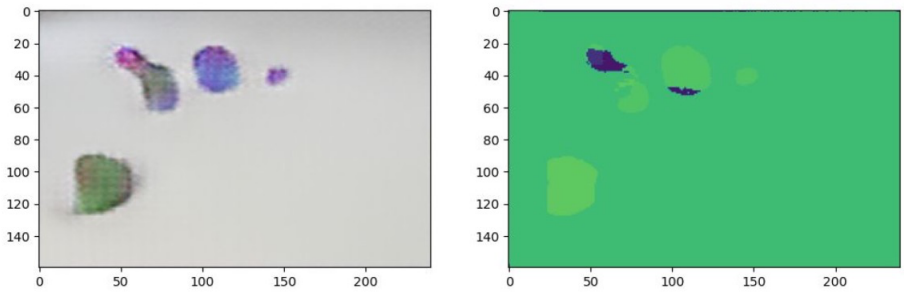


Figure 3: Prediction on FutureGAN Output by U-Net trained on 1k "train" videos

masks, offering better performance, especially in applications requiring high-fidelity visual data.

- **Exploring Alternative Video Prediction Algorithms:** Since the introduction of FutureGAN in 2018, numerous efficient video prediction algorithms have emerged, such as SimVP [5], a simple video prediction model completely built on CNN and trained using MSE loss in an end-to-end fashion. Future work can explore such algorithms, which have been shown to achieve state-of-the-art performance without additional tricks and complicated strategies.
- **Expanding Dataset Diversity:** To further improve the generalizability of our models, incorporating a more diverse set of training examples could be beneficial. This could include varying backgrounds, object types, and scenarios within the video frames, which would help the models learn more robust features and improve performance across a wider range of inputs.
- **Advanced Segmentation Techniques:** For U-Net, integrating techniques that address class imbalance and improve the segmentation of sparse classes could significantly enhance model performance. Techniques such as focal loss or SMOTE for augmenting training samples in under-represented classes might prove effective.

References

[1] Sandra Aigner and Marco Körner. Futuregan: Anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing autoencoder gans. *CoRR*, abs/1810.01325, 2018.

[2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[4] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *CoRR*, abs/1606.06650, 2016.

[5] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. Simvp: Simpler yet better video prediction, 2022.