Video Prediction and Mask Segmentation

Ashish Rai
New York University, CIMS
arr8134@nyu.edu

Harshit Mehta
New York University, CIMS
hm2540@nyu.edu

Jayesh Khandelwal
New York University, CIMS
jhk9774@nyu.edu

Abstract

This study addresses semantic segmentation in videos with limited labeled data, aiming to predict the 22nd frame mask using information from the initial 11 frames. Our proposed model integrates a transformer-based segmenter for semantic segmentation and ConvLSTM networks for video prediction. We explored trade-offs between deterministic and stochastic models, implementing MSPred. Code is available at: https://github.com/raishish/video-prediction-segmentation.

1 Introduction

In order to predict the 22nd frame mask from a sequence of only 11 historical frames is inherently challenging due to the stochastic nature of video data. In our scenario, objects exhibit diverse shapes (cube, sphere, and cylinder), materials (metal and rubber), and colors (gray, red, blue, green, brown, cyan, purple, and yellow). These objects undergo random motion and collisions, adhering to fundamental physics principles. Uncertainty arises from the unpredictable introduction of objects and their potential trajectories between frames 11 and 22. We possess an unlabeled dataset comprising 13,000 videos, each consisting of 22 frames. Additionally, our training and validation sets consist of 1,000 videos each, with each video comprising 22 frames and corresponding segmentation masks. To tackle this intricate task, we propose a two-step pipeline. First, we address the Video Prediction problem, aiming to forecast the next 11 frames given the initial 11 frames. Second, we delve into Mask Segmentation, focusing on accurately delineating object boundaries in the final frame. This dual approach allows for a comprehensive solution to the complex dynamics of object interactions and motion in the video sequence.

2 Related work

2.1 Video Prediction:

Earlier approaches relied on numerical and statistical methods, while recent advances integrate Convolutional Neural Networks (CNNs) for spatial feature extraction. Combining CNNs and Long Short-Term Memory (LSTM) networks, Convolutional LSTM networks excel in capturing spatial and temporal dependencies across diverse domains, offering improved forecasting accuracy and pattern recognition. Video prediction, essential for anticipative behavior, has evolved with deep-learning approaches. Oprea et al. [5] provide a comprehensive review, covering geometric transformations and recurrent networks with convolutional autoencoders. Stochastic models [1] introduce latent variables, while hierarchical structures like MSPred [8] leverage different abstraction levels. Focusing on simultaneous detailed and high-level frame prediction, MSPred utilizes RNNs with coarse temporal resolutions. Moreover, it extends the concept using convolutional LSTMs [6] and convolutional autoencoders for capturing distinct temporal resolutions in high-dimensional video sequences.

2.2 Segmentation:

Semantic segmentation's evolution from Fully Convolutional Networks (FCN) to transformer-based architectures is marked by approaches enhancing global context capture. Initial methods ([3], [4]) relied on convolutions and spatial pooling. Recent concerns about local operation limitations led to "Segmenter," a pure transformer architecture for semantic segmentation. Unlike convolution-based models, Segmenter utilizes transformers at every layer for global context integration. This shift aligns with the success of transformers in Natural Language Processing. Vision Transformer (ViT) [2] pioneered convolution-free image classification, and related works [9] extend transformers to video and semantic

segmentation tasks. Segmenter, featuring a ViT backbone [7], exemplifies this transition for competitive performance on segmentation benchmarks.

3 Proposed Model

The proposed novel model intricately addresses the task through a two-fold strategy:

3.1 Video Prediction

3.1.1 ConvLSTM-Based Video Prediction

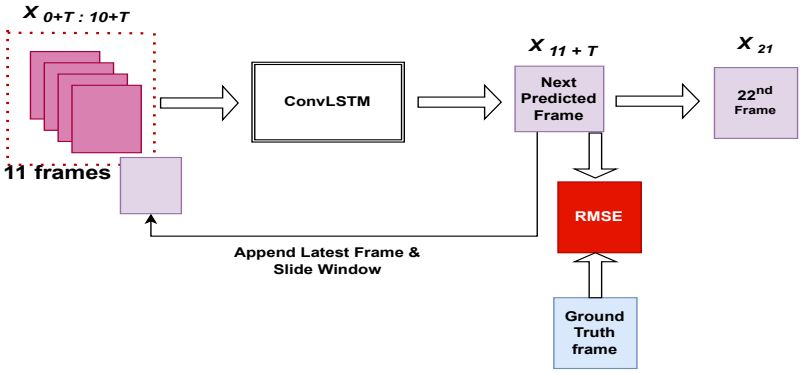


Figure 1: ConvLSTM

To capture the temporal evolution of video sequences, Convolutional LSTM networks were initially employed. Eleven stacked ConvLSTM cells, each housing 64 kernels, processed the input tensor. A subsequent CNN with three output channels predicted the succeeding frame. To counter the common issue of deterministic models producing blurry images due to their tendency to average outcomes, a shift was made towards stochastic modeling. This strategic transition introduced variability, enhancing the realism and diversity of frame predictions.

3.1.2 MSPred-Based Video Prediction

After an exhaustive literature review, MSPred emerged as the preferred stochastic model. Utilizing a structure akin to ConvLSTM, MSPred adopted a unique approach by modeling features as probability distributions in the latent space during training. Inference involved sampling latent variables, contributing to diverse frame predictions. The encoder and decoder, based on the VGG architecture, integrated CNN features into an LSTM model. The loss function incorporated both KL divergence and RMSE for improved performance. MSPred, akin to variational auto-encoders, adeptly samples latent variables given the historical 11 frames.

3.2 Mask Segmentation

In the second phase, a transformer-based encoder inspired by Vision Transformer (ViT) was employed for semantic segmentation. Operating on image frames, this encoder processed patches, transforming them into a latent space. This facilitated efficient spatial feature analysis within each patch. A mask transformer, leveraging latent patches, generated pixel-level segmentation masks. Evaluation metrics, including Intersection over Union (IOU), gauged the model's segmentation accuracy against ground truth masks. This comprehensive approach ensures robust video prediction and precise semantic segmentation.

4 Experimentation and Implemention

4.1 Dataset

The dataset comprised both training and validation sets, each consisting of 1,000 videos. Each video within these sets consisted of 22 frames accompanied by corresponding segmentation masks. Additionally, an unlabeled set was included, containing 13,000 videos, each comprising 22 frames. As, mention in section 1 we have 48 total object class, so including background we will have 49 classes.

4.2 Training

For **ConvLSTM**, two distinct training methods were applied using a sliding window on the training dataset.
Method_1: Leveraging the first 11 frames to predict the 12th frame, and subsequently incorporating target frames in the sliding window(SW) for subsequent predictions i.e. using 1-11 target frames to predict 12th frame, using 2-12 target frames to predict 13th frame, and so on.
Method_2: Similar to Method 1, but utilizing both the original frames and predicted frames within the sliding window(SW), i.e. using 1-10 target frames with 11th predicted frame to predict 12th frame. It was observed that employing predicted frames in the window sliding process yielded superior results.

Further refinement of the ConvLSTM model involved training on the unlabeled dataset to attain a minimal Root Mean Squared Error (RMSE) loss on the validation dataset. Integrating this model with the segmenter model resulted in an overall Intersection over Union (IOU) of 0.25. Notably, training the ConvLSTM (deterministic model) exhibited rapid convergence, achieving optimal results in under five epochs.

On the other hand, **MSPred** training involved utilizing both image frames and masks on the training set and solely image frames on the unlabeled set. Extensive experimentation revealed that simultaneous training on both image and mask frames led to early overfitting and unsatisfactory results, prompting rigorous hyperparameter tuning. The dimension of the latent space emerged as the most influential parameter, as the model transforms 3D image frames into this latent space, predicts within it, and then decodes the predicted latent space into image frames.

Turning to the **segmenter**, an abundant dataset of 22,000 image-mask pairs for both training and validation facilitated robust supervised learning. The model, optimized through meticulous hyperparameter tuning, achieved an impressive 88% Intersection over Union (IOU) on the segmentation task.

4.3 Results

Firstly for the segmenter, we meticulously fine-tuned various hyperparameters, as outlined in Table 1. Notably, we achieved an impressive Intersection over Union (IOU) of 88% for the specified parameters: { transformer embedding dimension (d_{emb}), transformer MLP head dimension (d_{mlp}), number of self-attention heads (h), number of encoder / decoder layers (l)}.

Before analyzing the video prediction results in Table 2, let's decipher the terminology: {"CL": ConvLSTM ; "MSP": MSPred; "seg": segmenter ; "SW": Sliding window method; "SW: PF": Sliding window method on predicted frames; "d_{lat}": dimesnion of latent variable; "β": Loss = (RMSE)^2 + β(KL Loss)} At the end of this project, our ConvLSTM + Segmenter model, employing the sliding window method on predicted frames during training, yields the highest mean IOU of **0.25** on the validation dataset.

Model	Hyperparameters				IOU
	d_{emb}	d_{mlp}	h	l	
seg	128	512	3	4	0.80
seg	128	756	4	4	0.85
seg	256	1024	4	6	**0.88**

Table 1: Segmenter Results

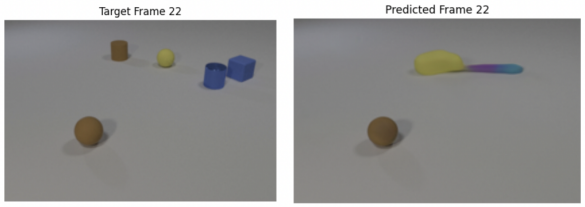
Model	Hyperparameters	RMSE	KL Loss	mIOU
CL (+ seg)	SW: Target frames	24	-	0.17
CL (+ seg)	SW: Predicted frames	12.99	-	0.22
CL (+ seg)	Unlabeled data (SW: PF)	12.63	-	**0.25**
MSP (+ seg)	$d_{lat} = 32$, $\beta = 0.001$	31.54	3546.23	0.13
MSP (+ seg)	$d_{lat} = 48$, $\beta = 0.001$	11.66	3189.5	**0.18**

Table 2: Video Prediction Results

In Figure 2, we present a comprehensive visualization of results obtained from ConvLSTM, MSPred, and Segmenter models. The ConvLSTM model, being deterministic, endeavors to ascertain the probability of an object within a frame. While it accurately predicts the location of stationary objects, its performance falters when it comes to future prediction for moving objects in subsequent frames. In such cases, it resorts to predicting an average of all potential outcomes, resulting in a blurred object representation. On the other hand, MSPred being a stochastic model, excels in recognizing object motion but tends to stretch the object, keeping the integrity of an object (a solid stays connected) insteading of breaking it into its components by generating blurred images. This not only increases the overall loss but also diminishes the Intersection over Union (IOU). Upon closer inspection, ConvLSTM emerges as the superior choice for predicting the 22nd frame.

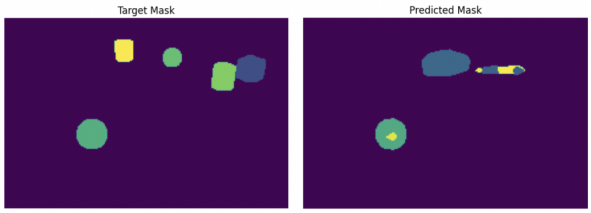
When it comes to segmentation masks, both ConvLSTM and MSPred employ the same model. In the case of ConvLSTM, when the object appears blurry, the segmenter predicts not only the real mask but also introduces inaccuracies by predicting additional masks. In contrast, MSPred accurately predicts the mask for a stretched object but suffers from mislocating the object, leading to a notable reduction in IOU.

(a) ConvLSTM: 22nd Target Frame vs Predicted Frame



(b) MSPred: 22nd Target Frame vs Predicted Frame

(c) ConvLSTM+Seg: 22nd Target Mask vs Predicted Mask



(d) MSPred+Seg: 22nd Target Mask vs Predicted Mask

Figure 2: Qualitative Result: ConvLSTM vs MSPred

5 Conclusion

In conclusion, our optimal model, ConvLSTM + Segmenter, demonstrates proficiency in predicting stationary objects and objects with limited movement. However, it encounters challenges when confronted with scenarios involving multiple objects in motion or collisions, leading to a blank prediction where only the background mask is generated. On the other hand, MSPred + Segmenter excels in predicting the future position or motion of an object but struggles to accurately preserve the object's shape.

Our experimentation has provided valuable insights into how models comprehend and address the task at hand. In future endeavors, we aim to enhance MSPred's performance by incorporating shape regularization. While our current approach involves utilizing $MSE + \beta * KL\ Loss$, our future work will explore the integration of Wasserstein distance in place of KL Loss, coupled with innovative shape regularization techniques. This ongoing exploration seeks to refine our models and further advance our understanding of complex tasks in video prediction.

Bibliography

[1] E. Denton and R. Fergus. Stochastic video generation with a learned prior, 2018.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[3] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, 2013. doi: 10.1109/TPAMI.2012.231.

[4] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation, 2015.

[5] S. Oprea, P. Martinez-Gonzalez, A. Garcia-Garcia, J. A. Castro-Vargas, S. Orts-Escolano, J. Garcia-Rodriguez, and A. Argyros. A review on deep learning techniques for video prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 2806–2826, 2022. doi: 10.1109/TPAMI.2020.3045007.

[6] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. kin Wong, and W. chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting, 2015.

[7] R. Strudel, R. G. Pinel, I. Laptev, and C. Schmid. Segmenter: Transformer for semantic segmentation. *CoRR*, abs/2105.05633, 2021.

[8] A. Villar-Corrales, A. Karapetyan, A. Boltres, and S. Behnke. Mspred: Video prediction at multiple spatio-temporal scales with hierarchical recurrent networks, 2022.

[9] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, 2021.

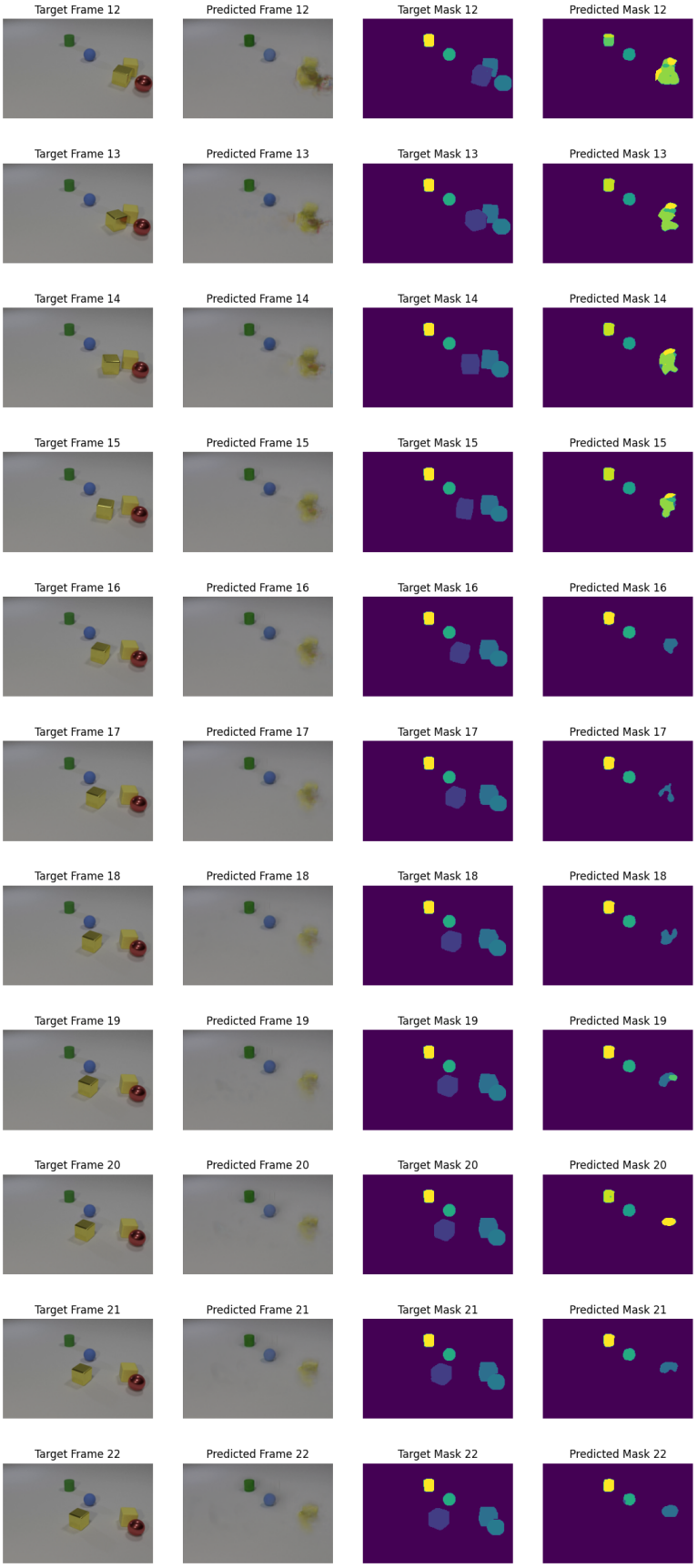


Figure 3: ConvLSTM + Segmenter Results (for all frames) on Validation dataset