Future Segmentation Prediction using SimVP and Semi-Supervision

Chanukya Vardhan Gujjula *
New York University
crg9968@nyu.edu

Pranav Jangir *
New York University
pj2251@nyu.edu

Sandeep Menon *
New York University
snm6477@nyu.edu

Abstract

In this work, we present various techniques for future segmentation prediction with limited labeled data. Our task is to predict the segmentation mask of the 22^{nd} frame, given the first 11 frames of synthetically generated videos of simple 3D objects that interact according to basic physics principles. We employ a deep learning model for spatiotemporal predictive learning called Simple Video Predictor (SimVP) and leverage semi-supervised learning to improve our predictions. We first train a deeplabv3 segmentation model on labeled data, and then adapt SimVP to predict segmentation masks directly. We further enhance our model by incorporating weak semi-supervision using the vast unlabeled data. To overcome the challenge of unpredictable objects that come in after the 11^{th} frame, we train our model on *clean* videos. We also introduce a decision tree based object suppression technique, named New Object Suppression (NOS), to correct the mistakes in model predictions. Our method achieves a Jaccard score of 44% on the validation set. Our findings demonstrate the potential of SimVP and semi-supervised learning in predicting future segmentation masks in synthetically generated videos.

1 Introduction

Convolutions are known for its ability to learn and extract spatial features and textures from images. By stacking such convolutional layers, we can learn more and more high level features, allowing us to learn tasks such as classification, detection, segmentation, etc., on images very effectively. However, when it comes to videos, which are a sequence of images, it was only natural to go into sequence models such as RNNs [8-10] and Transformers [11,12]. But from [1-7], we can see that even convolutions can learn temporal features when applied across the time dimension. In this paper we build upon the SimVP model from [1].

2 Related Work

Future segmentation prediction is closely related to video prediction and can be considered as a general spatiotemporal predictive learning task. Majority of existing approaches can be categorized into autogenerative or one-shot predictions that make use of only convolutional layers.

2.1 Auto Regressive

Following a general sequence-to-sequence framework, these approaches generate future predictions frame by frame using previous predicted frame for capturing temporal evolution. ConvLSTM[8] is an influential work in spatiotemporal predictive learning that extends a fully connected LSTM to use

[*] Equal contribution

Preprint. Under review.

convolutional structures in input and state transitions to predict future image frames. E3D-LSTM[9] uses separate 3D convolutions for encoding and decoding and integrates it into latent RNNs.

2.2 Fully Convolutional

Fully CNN based frameworks are less popular as they require complex modules and training strategies. [6] proposes a hierarchical architecture that makes predictions at different spatial resolution which are merged to generate future frames and train the model with adversarial and perceptual loss functions. [7] uses a 3D CNN forecasting module that is coupled with teacher-student distillation. SimVP[2] proposed a complete CNN based architecture trained on a simple MSE loss. It follows the common encoder, translator and decoder method with an inception-Unet translator in [2] and improved it further by using a gated spatiotemporal attention translator in [1] by simulating large kernel convolutions. This simple model has significant reduced training cost and makes it easier to scale to complex scenarios.

3 Method

As discussed in Section 1, we build upon the SimVP model[1], and in this section we focus on how we use SimVP and the training approaches followed.

3.1 Self-supervision

Self-supervision involves training a model on unlabeled data by defining a pretext task to learn underlying features that can be used for other downstream tasks. In our case, the data consists of 13K unlabeled videos and 1K labeled training and validation videos, with 22 frames each. We define our pretext task as predicting the future 11 frames given the first 11 frames and use the trained basic SimVP model on the downstream task of predicting the segmentation mask of the 22^{nd} frame.

3.2 Independent Segmentation

We trained a deeplabv3[13] segmentation model on the training set of images, and evaluated the performance our self-supervised model by segmenting the predicted 22^{nd} frame with ground-truth label, resulting in a jaccard of 22.2% on the validation set. This baseline indicates there is room for improvement in our self-supervised model.

3.3 Segmentation using SimVP

Since SimVP is able to predict the raw future frames, the model should have learned the underlying features to generate color, shape and texture of objects during self-supervision. It seemed redundant to have a separate segmentation network predict the mask after constructing the raw image, as these learned features by the encoder and translator could be used to generate this mask.

We reset the decoder of SimVP with fresh weights, and modified the final convolutional layer to output 49 channels (number of segmentation classes) instead of RGB channels. This modified SimVP was trained on the training set of videos with a higher learning rate for the modified decoder and a low learning rate for the encoder and translator. This allowed the decoder to learn to predict the mask and fine-tuned the weights of encoder and translator to this task. This resulted in a jaccard of 31.4%, indicating that the modified SimVP was able to generate more accurate segmentation masks.

3.4 Weak semi-supervision

Although the modified SimVP resulted in better predictions, we are only able to train on the small set of 1K labeled train videos. In order to make use of the vast unlabeled videos, we decided to leverage our trained deeplabv3 model to generate labels for this unlabeled videos. Despite the fact that the our deeplabv3 model was not perfect (94.1% validation accuracy), this weak-semi supervision step was shown to improve the performance[15]. By training the modified SimVP on the weak segmentation labels, we were able to achieve a jaccard of 38.47%.

3.5 Data Cleaning

Given the task of predicting the 22^{nd} frame using only the first 11 frames, we believed it was impossible for a model to account for new objects appearing after the 11^{th} frame. The presence of such videos in the dataset confused the learning process as the model tried to predict unseen objects, affecting predictions for known objects. However, the SimVP model should be able to predict trajectories for known objects without interference from new ones.

To improve the model's performance, we identified videos with no new objects appearing after the 11^{th} frame and fine-tuned the model on that subset. This led to an increased validation Jaccard Score of 42.14%.

3.6 New Object Suppression

Even after training on clean data, there is a small change that the model still predicts spurious new objects. To clean up predicted masks in such cases, we designed a decision tree, as shown in Figure . A known object is defined as an object with a class that was present in the first 11 frames.

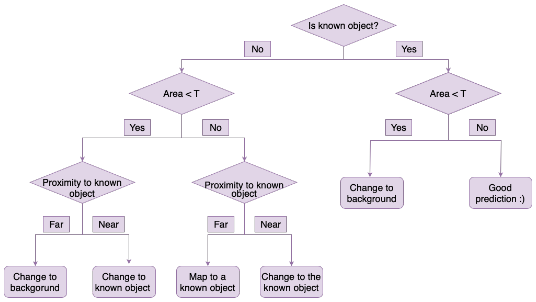

Figure 1: Decision tree for New Object Suppression

The logic can be summarized as follows:

- If it is not a known object and is near to a known object, map it to the known object.
- If it is not a known object and is far from a known object, change it to background if area is less else map it to a known object (based on the difference in material, shape and color in the same priority).
- For a known object, if the area is too small it is considered noise and is changed to the background. If the area is not small, it is considered as a good prediction.

Applying NOS on modified SimVP trained on clean data further pushed the jaccard score to 44.17%.

4 Results

In our study, we aimed to predict the next 11 frames of a video using a SimVP model. We then fine-tuned our model to predict masks instead of RGB images, which led us to change our loss metric from MSE to Jaccard. On the validation set, we obtained a Jaccard score of 0.42. To further improve our results, we applied decision tree heuristics, which resulted in a Jaccard score increase from 0.42 to 0.44 on the validation set. Find examples in Figures 2 and 3.

5 Future Work

The self-supervision and weak semi-supervision approaches discussed through the paper make use of SimVP to construct the future raw image frames or the corresponding masks. However, we can follow JEPA[14] based training where we directly learn in the embedding space through a reconstruction loss. Specifically, we can use the spatial encoder of SimVP to encode both input frames and future frames, and apply a reconstruction loss between the embedding of future frames and the output of translator that takes in embedding of input frames (see Figure 4). We believe this would enable the

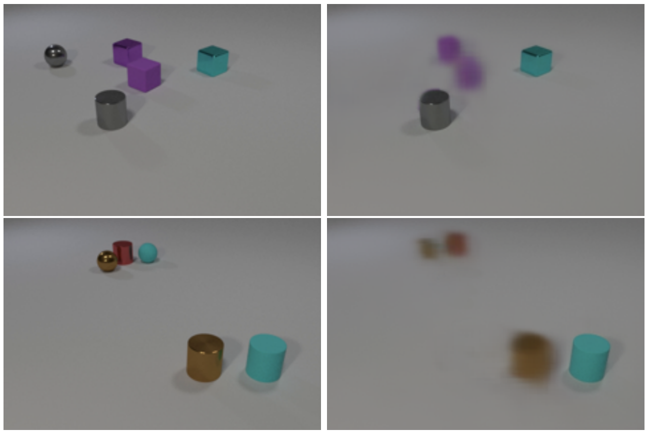

Figure 2: Two examples from video number 3 (top row) and 4 (bottom row). True (left) and predicted (right) 22^{nd} frame.
(Video 3) We can see that the metal grey ball is present in the true frame and not in the predicted frame. This is because this ball enters the video in the 17^{th} frame and no model can predict this. This shows that our model is learning collision physics conditioned on interactions between objects that the model has seen till the 11^{th} frame.
(Video 4) The cyan rubber cylinder is predicted perfectly whereas the other objects are blurry. The cyan rubber ball is absent from the predicted frame, which is a flaw of the our model as all the objects were present in the first 11 frames. According to our predictions, the objects are a bit far from where they should be and our model (wrongly) predicts that the cyan rubber ball has moved out of the frame.

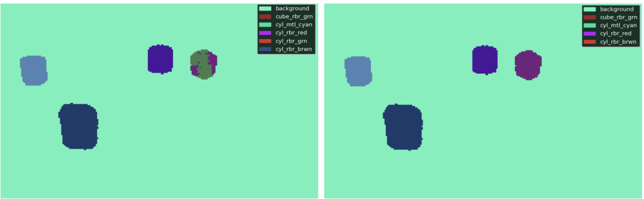

Figure 3: Predicted 22^{nd} frame without (left) and with (right) heuristics applied. In the left mask, the violet object is a known object and the green objected detected within it is a new object. Our heuristics fix this by looking for the closest known object and replacing the green one.

model to learn a better energy manifold with the 22 frames in the low energy valley. Moreover, this could further be used in downstream tasks, such as mask segmentation of visual question answering.

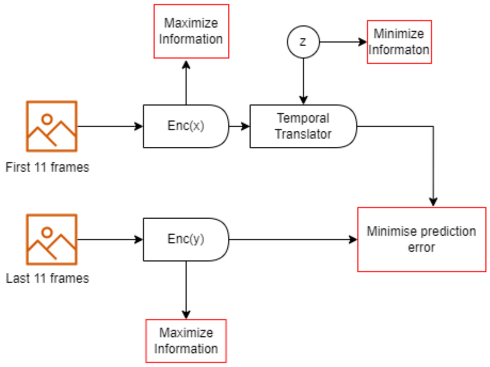

Figure 4: SimVP model in a Joint Embedding Predictive Architecture (JEPA)

References

[1] C. Tan, Z. Gao, S. Li, and S. Z. Li, "Simvp: Towards simple yet powerful spatiotemporal predictive learning." arXiv preprint arXiv:2211.12509, 2022.

[2] Z. Gao, C. Tan, L. Wu, and S. Z. Li, "Simvp: Simpler yet better video prediction." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3170-3180, 2022.

[3] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error." arXiv preprint arXiv:1511.05440, 2015.

[4] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow." In Proceedings of the IEEE International Conference on Computer Vision, pp. 4463–4471, 2017.

[5] M. Henaff, J. Zhao, and Y. LeCun, "Prediction under uncertainty with error-encoding networks." arXiv preprint arXiv:1711.04994, 2017.

[6] O. Shouno, "Photo-realistic video prediction on natural videos of largely changing frames." arXiv preprint arXiv:2003.08635, 2020.

[7] H. K. Chiu, E. Adeli, and J. C. Niebles, "Segmenting the future." In IEEE Robotics and Automation Letters, vol. 5, no. 3, pp. 4202–4209, 2020.

[8] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting." Advances in Neural Information Processing Systems, vol. 28, 2015.

[9] Y. Wang, L. Jiang, M. H. Yang, L. J. Li, M. Long, and L. Fei-Fei, "Eidetic 3d lstm: A model for video prediction and beyond." In International conference on learning representations, 2018.

[10] M. Babaeizadeh, M. T. Saffar, S. Nair, S. Levine, C. Finn, and D. Erhan, "Fitvid: Overfitting in pixel-level video prediction." arXiv preprint arXiv:2106.13195, 2021.

[11] D. Weissenborn, O. Tackstrom, and J. Uszkoreit, "Scaling autoregressive video models." arXiv preprint arXiv:1906.02634, 2019.

[12] R. Rakhimov, D. Volkhonskiy, A. Artemov, D. Zorin, and E. Burnaev, "Latent video transformer." arXiv preprint arXiv:2006.10704, 2020.

[13] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam. "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv:1706.05587, 2017.

[14] Y. LeCun, "A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27.", 2022.

[15] Li, Qizhu, Anurag Arnab, and Philip HS Torr. "Weakly-and semi-supervised panoptic segmentation." Proceedings of the European conference on computer vision (ECCV). 2018.