# FUTURE SEGMENTATION PREDICTION USING SIMVP AND SEGFORMER

A PREPRINT

**Anirudh Garg** **Nikhil Kommineni** **Sai Preetham**

January 6, 2024

**ABSTRACT**

In this work, we present different solutions for future segmentation prediction. Our goal is to predict the 22nd frame using the first 11 frames in synthetic videos consisting of 48 different objects distinguished by shape, color and texture. We employ CNN based models for motion prediction and transformer based models for segmentation mask computations. We first train SimVP to predict the 22nd frame from the first 11 frames and use this model to predict the 22nd frame as an intermediate result. We then train the Segformer model to predict segmentation masks of images and run it on the frames predicted by SimVP to get the final output. We use the Jaccard similarity index to evaluate our model.

***Keywords*** Video prediction · Optical flow prediction · Image Segmentation

## 1 Introduction

Video frame prediction and Semantic Segmentation are popular computer vision tasks with applications in autonomous-driving, path-planning and video compression. With the inherent complexity and randomness of video, lots of novel deep-learning models have been introduced in the recent years. Nueral network architectures ranging from RNNs and transformers to more sophisticated architectures like auto-regression started gaining popularity. In contrast, SimVP was introduced as fully CNN based architecture that achieved state-of-the-art performance on five datasets. We heavily-utilize the SimVP architecture to learn the spatio-temporal features of our dataset, we move forward to predict the segmentation masks of the final frames using Segformer, a transformer based architecture for semantic segmentation on images.

## 2 Background

Future Segmentation Prediction is heavily based on Future Frame Prediction or Deep Video Prediction.

Another component of the problem is the segmentation of the predicted video frame.

**Problem Statement**[1] Video frame prediction aims to infer future frames from the previous ones. Given a video sequence $\mathbf{X}_{t,T} = \{x_i\}_{t-T+1}^{t}$ at time $t$ with the past $\mathbf{T}$ frames, the goal is to predict the future sequence $\mathbf{Y}_{t,T'} = \{x_i\}_{t}^{t+T'-1}$ at time $t$ that contains the next $\mathbf{T}'$ frames, where $x_i \in \mathbb{R}^{C\times H\times W}$ is an image with $C$ channels, height $H$ and width $W$. Formally, the prediction model is mapping $\mathcal{F}_\Theta : \mathbf{X}_{t,T} \rightarrow \mathbf{Y}_{t,T'}$, with learnable parameters $\Theta$ optimized by,

$$\Theta^* = argmin_\Theta \mathcal{L}(\mathcal{F}_\Theta(\mathbf{X}_{t,T}), \mathbf{Y}_{t,T'})$$

where $\mathcal{L}$ is a loss function.

We solve this prediction problem for $\mathbf{T} = \mathbf{T}' = 11$ on a synthetic moving objects dataset.

**Segmentation** Image segmentation is a computer vision task that involves dividing an image into meaningful and

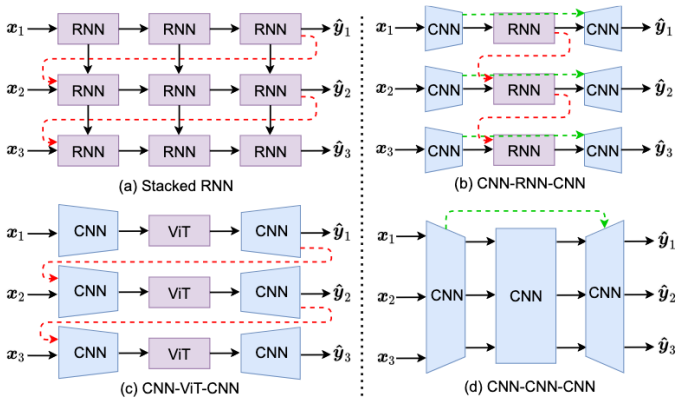


Figure 1: Different architectures for video prediction

distinct regions or segments based on certain criteria. The goal of image segmentation is to partition an image into semantically meaningful parts, making it easier to analyze and understand. The primary objective is to group pixels or regions in an image that share similar characteristics, such as color, intensity, texture, or other visual properties.

### 2.1 Related Work

Future frame prediction is a popular research problem. Optical flow estimation techniques were used to approximate the motion of every pixel to predict the frame at a future time[2]. A flow-vector is learned which is used to predict the future position of each pixel. In contrast to predicting the motion of each pixel CNN based models were suggested for frame prediction. Following the work on CNN architectures, LSTM and RNN based architectures were introduced[3, 4]. X. Shi et.al[4] discovered that ConvLSTMs had better performance compare to full-connected LSTMs in precipitation nowcasting.

Zhangyang Gao et.al [1], further classied deep-learning based architecture into four categories in Figure 3, i) RNN-RNN-RNN, ii) RNN-CNN-RNN, iii) CNN-ViT-CNN and iv) CNN-CNN-CNN based on the building blocks used in the models. RNN and ViT based models were primarily using CNN blocks to compress the input data and the corresponding RNN/ViT blocks to learn the spatio-temporal dynamics in the frame sequence.

Recent works in semantic segmentation have seen advancements in Transformer-based architectures, such as ViT [5] and its derivatives, showcasing state-of-the-art performance. These architectures, originally designed for image classification, have been adapted to pixel-level tasks, demonstrating effectiveness in dense prediction. Notable approaches include SETR [6], which adopts ViT for semantic segmentation, and PVT [7], introducing a pyramid structure in Transformers for dense prediction tasks. While Transformer-based methods show promise, their computational demands pose challenges for real-time applications. Prior research has also explored Transformer applications in various computer vision tasks beyond segmentation, including object detection, tracking, and multi-modal learning.

## 3 Our Approach

Given SimVP is a completely CNN based architecture that achieves state of the art perfomances on Moving MNIST, Human3.6 and KITTI [1], we chose to combine it with the Segformer model to predict future segmentation masks. This two-step approach that first predicts the 22nd frame for a video and then computes the segmentation mask on the predicted frame.

### 3.1 Frame Prediction

We are using SimVP to predict the 22nd frame using the first 11 frames of the video. SimVP[1] consists of an encoder block, a translator block and a decoder block, each of which uses convolutional layers. The encoder is used to learn spatial features, the translator is used to learn temporal features and the decoder is used to combine the spatial and temporal features to predict the future frames. Figure 2 captures a detailed architecture of SimVP.

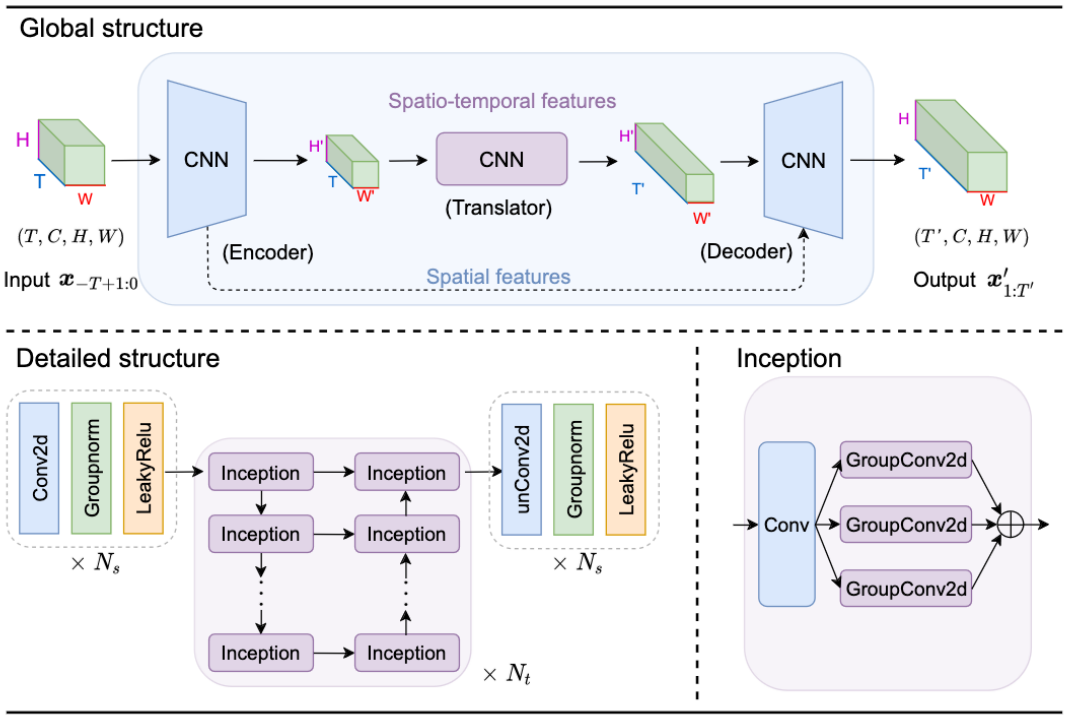


Figure 2: SimVP architecture[1]. The encoder consists of $N_s = 2$ ConvNormReLU blocks (Conv2d+LayerNorm+LeakyReLU, the decoder consists of $N_t = 8$ instances of bottleneckConv2d with 1×1 kernel followed by parallel GroupConv2d with each instance stacked on top of the previous one. The decoder utilizes $N_s$ unConvNormReLU blocks(ConvTranspose2d+GroupNorm+LeakyReLU)

We modified the original SimVP architecture to increase the size of the hidden layers from 256 to 512 observing gains in the jaccard index.

### 3.2 Segmentation

The proposed SegFormer framework [8] consists of two main modules: A hierarchical Transformer encoder to extract coarse and fine features; and a lightweight All-MLP decoder to directly fuse these multi-level features and predict the semantic segmentation mask."FFN" indicates feed-forward network. The encoder in the architecture processes input patch embeddings through four transformer blocks, incorporating Overlap Patch Merging, Self-Attention Layers, Skip Connections, and Normalization Layers. This hierarchical structure produces multi-scale features. Subsequently, the decoder receives these features from the encoder as input, utilizing an MLP-based architecture that incorporates Upsampling and Conv2D Layers. The final output of the decoder is a predicted segmentation mask, representing the model's semantic segmentation of the input data. This design aims to capture intricate patterns and contextual information through the hierarchical processing of input patches, providing an effective framework for image segmentation tasks.

### 3.3 Experiments

We train the SimVP based frame prediction model on an unlabeled dataset of 13,000 videos, each consisting of 22 frames of size 160×240×3, we use a hidden dataset of the first 11 frames to predict the 22nd frame. The 22nd frame predictions are then passed to the Segformer based semantic segmentation model that is trained on 1,000 labelled videos of 22 frames each to predict the mask for the final frame. The segmentation output consists of a mask that classifies 49 different combinations of object characteristics for each of the 11-frame videos.

The frame prediction model was trained for 30 epochs on the unlabelled data using vanilla MSE loss. We used The Adam optimizer for hyperparameter tuning. The Segformer model was trained for 20 epochs on the labelled data using cross entropy loss. We used a learning rate of 0.0001 and an Adam Optimizer.

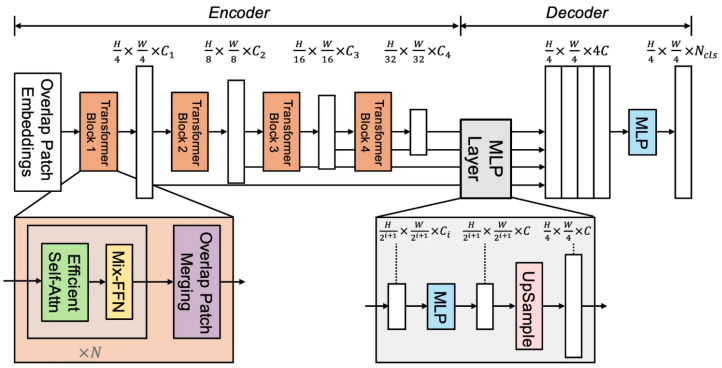


Figure 3: Segformer Framework [8]

Table 1: Segformer: Jaccard Index

| Epochs | Jaccard Index |
|---|---|
| 5 | 0.834 |
| 10 | 0.9435 |
| 20 | 0.9712 |

## 4 Results

We evaluated Segformers model's performance on 1,000 images from the validation dataset. The Jaccard Index values obtained from this experiment are presented in Table 1. The Segformer model trained on 20 epochs performed the best on these unseen images, achieving a maximum Jaccard Index score of 0.9712. This suggests that our segmentation model has the ability to predict well. The Jaccard Index values were obtained with our combined pipleine of SimVP + Segformer predictions on the entire validation dataset of 1,000 video clips and are presented in Table 2. The best performing model achieved a Jaccard Index of 0.2213 and was obtained from the combination of 25 epochs of SimVP training with a learning rate of 0.001. Figure4, 4 and 4 shows how features are learnt over the epochs. The final jaccard index achieved on the hidden dataset is 0.2441 based on the final results.

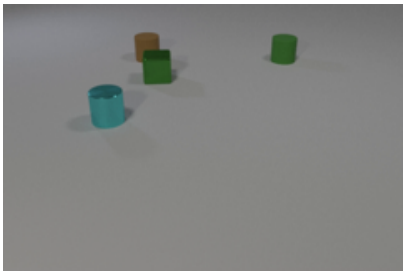

Figure 4: Actual 22nd frame

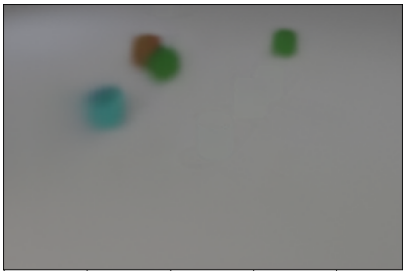

Figure 5: Predicted 22nd frame

Figure 6: Predicted 22nd mask

## References

[1] Zhangyang Gao and Cheng Tan and Lirong Wu and Stan Z. Li. SimVP: Simpler yet Better Video Prediction. In *arXiv preprint arXiv:2206.05099*, 2022.

[2] Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *arXiv preprint arXiv:2003.12039*, 2022

[3] Yunbo Wang and Zhifeng Gao and Mingsheng Long and Jianmin Wang and Philip S. Yu. PredRNN++: Towards A Resolution of the Deep-in-Time Dilemma in Spatiotemporal Predictive Learning. *arXiv preprint arXiv:1804.06300*, 2018.

Table 2: SimVP + Segformer: Jaccard Index

| Epochs | Jaccard Index |
|---|---|
| 10 | 0.00529 |
| 20 | 0.1332 |
| 30 | 0.2441 |

[4] Xingjian Shi and Zhourong Chen and Hao Wang and Dit-Yan Yeung and Wai-kin Wong and Wang-chun Woo. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *arXiv preprint arXiv:1506.04214*, 2015.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby : An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale In *arXiv preprint arXiv:2010.11929*, 2020.

[6] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, Li Zhang : Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers In *arXiv preprint arXiv:2012.15840*, 2021.

[7] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, Ling Shao : Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions In *arXiv preprint arXiv:2102.12122*, 2021.

[8] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo : SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers In *arXiv preprint arXiv:2105.15203*, 2023.